Clasificador de máquinas de vectores de soporte para problemas desbalanceados con selección automática de parámetros

Este trabajo de grado es presentado como una requerimiento para obtener el título de:
Magister en Ingeniería Eléctrica

Cristian Alfonso Jiménez Castaño

Tutor: Andrés Marino Álvarez Meza - MEng, PhD

Universidad Tecnológica de Pereira
Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la Computación
Maestría en Ingeniería Eléctrica - Línea de automática
Grupo de Investigación Automática
Pereira-Risaralda
2018

Índice general

Índice de figuras

Índice de tablas

1. Agradecimientos

Quiero agradecer a mi familia por su continuo e incondicional apoyo durante toda mi vida, en especial a mis padres, que nunca han dejado de enseñarme. También en especial a mi tio, mi segundo padre, José Nelson por ese apoyo incondicional a mi y a toda la familia, por demostrarnos que con trabajo arduo, dedicación y motivación todo es posible. A mis Abuelas y Abuelo por ser esas personas incondicionales en mi vida, que con amor y compresión nos han ayudado a pararnos y seguir adelante. También, quiero agradecerle a mi tutor Andrés Marino Álvarez M, por sus contribuciones a mi trabajo, a mi formación como persona y profesional, y en especial por esa calidad humana que posee. Estoy más que agradecido con la Maestría en Ingeniería Eléctrica del programa de Ingeniería Eléctrica en la Universidad Técnologica de Pereira y con el grupo de investigación en Automática.

Esta investigación es desarrollada bajo el proyecto "Desarrollo de un sistema de soporte clínico basado en el procesamiento estocástico para mejorar la resolución espacial de la resonancia magnética estructural y de difusión con aplicación al procedimiento de la ablación de tumores", code: 111074455860, financiado por COLCIENCIAS. Además este trabajo fue parcialmente financiado por el proyecto E6-18-09: "Clasificador de máquinas de vectores de soporte para problemas desbalanceados con selección automática de parámetros", de la Vicerrectoria de Investigación, innovación y extensión de la Universidad Tecnológica de Pereira.

2. Resumen

La mayoría de los métodos de clasificación asumen que el número de muestras en las clases estudiadas son las mismas (balanceadas). Sin embargo, realizar esta asunción puede llevar a desempeños sesgados, ya que, la mayoría de aplicaciones y bases de datos reales no son balanceadas, llevando a que estos métodos ignoren la clase minoritaria (la clase con el menor número de muestras). Este trabajo propone un clasificador novedoso, llamado *enhanced twin support vector machine*–(ETWSVM), que representa las muestras de entrada en un espacio de características de alta dimensionalidad, posiblemente infinita, durante la construcción de una frontera de decisión bajo la filosofía del *twin support vector machine*–(TWSVM). También, usamos un método basado en *centered kernel alignment*–(CKA) para aprender la función kernel con el fin de contrarrestar los problemas inherentes del desbalance y mejorar la separabilidad de los datos. Además, adoptamos las estrategias *One-versus-Rest* y *One-versus-One* para extender la formulación del ETWSVM a tareas de clasificación multiclase. De los resultados obtenidos sobre bases de datos sintéticas y reales, nuestra propuesta supera métodos del estado del arte con respecto al desempeño (precisión, media geométrica, F-measure), y tiempo de entrenamiento. En efecto, después analizamos la sensibilidad de los parámetros libres para diferentes tasas de desbalance y traslape entre las clases, y sugerimos una variante del ETWSVMN automático que registra una indicada relación entre desempeño de clasificación y tiempo de entrenamiento.

3. Abstract

Most of the classification approaches are based on the assumption that the sample distribution among classes is balanced. Nonetheless, such an assumption leads to biased performance over the majority class (the class with the highest number of samples). This work proposes a novel approach, called *enhanced twin support vector machine*–(ETWSVM), that represents the input samples in a high-dimensional feature space of possible infinity dimension during the decision boundary building under a *twin support vector machine*–TWSVM philosophy. Also, we use a *centered kernel alignment*–(CKA)-based approach to learning the kernel function counteracting inherent imbalanced issues and improving the data separability. Moreover, we also adopt One-vs-One and One-vs-Rest frameworks to extend the ETWSVM formulation for multi-class tasks. Obtained results on synthetic and real-world datasets show that our approach outperforms state-of-the-art methods concerning classification performance (accuracy, geometric mean, and F-measure), and training time. Indeed, after analyzing the sensitive of the free parameters for the imbalanced ratio and the overlapping among clusters, we suggest an automatic ETWSVM variant that achieves a suitable trade-off between classification performance and training time.

4. Notación

Símbolo	**Descripción**
$\mathbb{R}^P$	espacio Euclidiano de dimensión P
a	los escalares se representaran por letras minúsculas
P	los escalares que indican dimensiones y tamaño de vectores y/o matrices se expresarán en mayúscula y sin negrilla. En especial, este indica la dimensión del espacio de entrada de los datos
P'	dimensión del espacio embebido por selección o extracción de características
Q	dimensión del espacio de características, RKHS, $\mathcal{H}$
N_+	número muestras de entrada pertenecientes a la clase minoritaria
N_-	número muestras de entrada pertenecientes a la clase mayoritaria
N	número total de muestras de entradas $N{=}N_+ + N_-$
$N_{\mathcal{S}}$	Número de vectores de soporte total del modelo entrenado
ℓ	índice que hace referencia a la clase mayoritaria (-1) o minoritaria ($+1$), por facilidad este toma valores del conjunto $\ell{\in}\{+,-\}$. Su reciproco es $\ell'{=}{-}\ell$
$\boldsymbol{x}$	los vectores columnas serán representados en letras minúsculas y en negrilla. Se asumirá que estos pertenecen a $\mathbb{R}^P$ siempre y cuando no se especifique lo contrario
$\boldsymbol{A}$	las matrices serán indicadas con letras mayúsculas y en negrilla
$a_{nn'}$	elemento de la fila n y columna n' de la matriz $\boldsymbol{A}$
$\boldsymbol{\alpha}_{\ell'}$	vector de los multiplicadores de Lagrange de las muestras de la clase $\ell' 1$, donde $\ell'{\in}\{+,-\}$
$\boldsymbol{I}$	matriz identidad, una matriz cuadrada de dimensiones apropiadas, tiene 1's en su diagonal y 0's fuera de esta
$\|\cdot\|_2$	norma Euclidiana o también conocida como norma L_2
$\|\cdot\|_{\mathcal{H}}$	norma Euclidiana o también conocida como norma L_2 en el espacio RKHS $\mathcal{H}$
$\|\cdot\|$	valor absoluto de $\cdot$
$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_{\mathcal{H}}$	producto interior entre los vectores columna $\boldsymbol{x}_i$ y $\boldsymbol{x}_j$ in el espacio $\mathcal{H}$, donde $\boldsymbol{x}_i, \boldsymbol{x}_j{\in}\mathcal{H}$
$\mathbf{1}_\ell$	vector de longitud N_ℓ con sólo valores de 1's
$\mathbf{0}_M$	vector de longitud M con sólo valores de 0
$f(\cdot)$	función escalar evaluada en $\cdot$
sign$(\cdot)$	función *signo* evaluada en $\cdot$
med$(\cdot)$	operador mediana de $\cdot$
$\varphi(\cdot)$	función de mapeo desde el espacio de entrada $\mathcal{X}$ al espacio de características, o RKHS, $\mathcal{H}$

5. Introducción

5.1. Motivación

En la formulación de métodos de aprendizaje para clasificación una de las primeras asunciones es que el número de muestras entre las clases estudiadas son iguales, es decir que se va aplicar sobre bases de datos balanceadas. El problema de desbalance de datos en clasificación ha captado, en los últimos años, un gran interés por parte de la comunidad científica. Debido, a que varias aplicaciones del mundo real tienen muchas muestras de un fenómeno en comparación al número de muestras de otro, algunos ejemplos de esto se pueden encontrar en: la detección de intrusión [1], el diagnóstico médico [2], el diagnóstico de cáncer [3], el análisis de tráfico [4], y valoración de crédito [5].

El principal problema del desbalance de datos es la gran posibilidad de que la clase minoritaria, la clase con el menor número de muestras, sea sesgada por el clasificador. Debido a que esta clase puede ser tratada como ruido, o que el ruido puedan ser tratados como de esta clase, por el clasificador [6].

5.2. Planteamiento del problema

En la actualidad este problema es atacado principalmente desde tres enfoques. El primero es a nivel de los datos en el cual se realiza un pre-procesamiento de los mismos, el más común son los métodos de remuestreo en los cuales se busca balancear el número de muestras entre las clases por medio de la eliminación y/p generación de muestras de las clases [7]. El segundo enfoque es el desarrollo de algoritmos costo-sensible, donde se pondera la función de costo del algoritmo de aprendizaje de tal forma que la mala clasificación de la clase minoritaria se penalice con mayor fuerza [8]. El tercer enfoque es el ensamblaje de clasificadores, donde se fragmenta la clase mayoritaria en subconjuntos de la misma, sin que se sobrepongan, y son entrenados tantos clasificadores como subconjuntos obtenidos, la predicción se realiza por votación mayoritaria [9].

Entendiendo un poco lo anterior. El remuestreo se puede dividir en tres categorías: submuestreo de la clase mayoritaria, sobremuestreo de la clase minoritaria y el híbrido entre ellos. El submuestreo, consiste en extraer muestras de la clase mayoritaria para así balancear el número de muestras entre ambas clases, el principal problema del submuestreo es la pérdida potencial de información debido a la eliminación de las muestras procedentes de la clase mayoritaria. Habría que decir también, que el sobremuestreo consiste en expandir el conjunto de muestras de la clase minoritaria por medio de la creación de muestras sintéticas, a partir de la originales; pero este tiende a generar sobre-entrenamiento por la información redundante procedente de estas muestras, además, el tiempo de cómputo crece drásticamente debido a la creación de dichas muestras. Cabe destacar, que los métodos de remuestreo son indiferentes al clasificador a usar [7]. El segundo enfoque, ensamblaje de clasificadores, consiste en fragmentar el conjunto de muestras mayoritarias en subconjuntos de las mismas, de tal forma que no exista intersección entre ellas, del mismo tamaño de la clase minoritaria. Luego, por cada subconjunto de la clase mayoritaria y el conjunto de la clase minorita-

ria se entrena un clasificador, al final se entrenan tantos clasificadores como subconjuntos se extraigan.Por lo general las metodologías basadas en esta filosofía presentan un buen rendimiento pero su costo computacional es alto, por el número de clasificadores a entrenar [10]. Por último, el enfoque por algoritmos de costo-sensible pueden llevar fácilmente a sobre-entrenamiento [8].

En años recientes, han derivado métodos basados en máquinas de vectores de soporte (*support vector machine*–(SVM)) con el fin de replicar su capacidad de generalización con un costo computacional menor [11]. Uno de estos métodos son las máquinas de vectores de soporte gemelas (*twin support vector machine*–(TWSVM)) las cual soluciona dos problemas de programación cuadráticas (QPPs) más pequeñas que el QPP que soluciona la SVM. Lo anterior, lleva a que las TWSVM tengan un costo computacional menor que las SVM con una buena capacidad de generalización [12]. No obstante, poseen el mismo problema ante bases de datos desbalanceadas que la SVM estándar. Algunas extensiones del TWSVM incluyen una representación de la estructura de los datos para favorecer su capacidad de generalización de la clasificación [13]. Adicionalmente, han sido propuestos algoritmos basados en TWSVM para contrarrestar los problemas de desbalance de datos, combinándolo con técnicas de re-muestreo, además de problemas de optimización ponderados con el fin de evitar sesgo en la discriminación [14–16]. Aún así, los enfoques basados en TWSVM no incluyen una representación apropiada a un espacio de Hilbert con kernel reproductivo (RKHS) por medio una función de mapeo, la cual mapea las muestras del espacio original a un espacio de alta dimensionalidad que mejore la representación de los datos. Por lo tanto, descuidan la formulación intrínseca y las virtudes del problema dual con respecto a los métodos kernels.

De lo anterior se puede concluir que, los principales problemas del desbalance de datos son: tiempo de cómputo, sobre-entrenamiento y pérdida de información. También se puede concluir que el problema de desbalance no se ha podido solucionar por completo y nos deja con la pregunta de investigación: ¿Es posible el desarrollo e implementación de una metodología basada en TWSVM con una nueva extensión no-lineal con mapeo a un espacio de alta dimensionalidad, posiblemente infinito, el cual contrarreste los efectos del desbalance en problemas de clasificación?, además ¿es posible la selección automática de los parámetros libres en aras de reducir el tiempo de entrenamiento y favorecer la identificación de patrones relevantes?.

5.3. Estado del Arte

En el estado del arte las técnicas de remuestreo se dividen en: submuestreo de la clase mayoritaria, sobremuestreo de la clase minoritaria e híbrido. La técnica de submuestreo más eficiente y simple es el submuestreo aleatorio (RUS, por sus siglas en inglés), el cual consiste en la eliminación aleatoria de muestras de la clase mayoritaria [17]. Dos métodos ampliamente utilizados como técnicas de sobremuestreo son duplicación aleatoria de muestras minoritarias y el *synthetic minority over-sampling technique*–(SMOTE) [18]. [19] realizó un estudio de los tipos de clases minoritarias y su influencia en el aprendizaje del clasificador en datos desbalanceados. Algunos autores han tratado de definir de forma automática la proporción de muestreo para diferentes tasas de desbalance y configuración del problema [20–25]. Por otra parte, [7, 26] han realizado un estudio sobre el desempeño de diferentes técnicas de remuestreo, de los cuales se puede concluir: que cuando se tiene cientos de muestras de la clase minoritaria, el submuestreo supera al sobremuestreo en términos de tiempo de computó; cuando se tienen pocas muestras de la clase minoritaria, la mejor opción es aplicar alguna técnica de sobremuestreo; si el tamaño de las muestras de entrenamiento es muy grande, lo mejor es combinar ambos tipos de técnicas de remuestreo; para finalizar, los métodos de sobre-muestreo son un poco más eficaces para reconocer los datos atípicos.

Por otra parte, la estrategia de ensamblamiento de clasificadores es muy popular pero conlleva un gran costo

computacional para entrenar varios clasificadores. Luego, esta estrategia se divide principalmente en dos enfoque: Bagging y Boosting. El Bagging consiste en que los clasificadores base pueden ser entrenados en paralelos. Por otro lado, el Boosting consiste en un algoritmo de entrenamiento de los clasificadores bases, el cual es evolutivo y calcula los pesos que van a tener cada clasificador base en la predicción. Este último enfoque es el más común y eficiente, el primer método de esta naturaleza fue AdaBoost propuesto en 1996 [27]; en el estado del arte se encuentra una gran variedad de métodos basado en este y con diferentes tipos de clasificadores base, algunos ejemplos son: [9, 10, 28, 29].

Por último, los algoritmos de costo-sensibles comparados con los métodos de remuestreo, los algoritmos de aprendizaje son más eficientes computacionalmente hablando, por lo que puede ser más adecuado para grandes flujos de datos. Sin embargo, es más popular los métodos de remuestreo para escenarios de desbalance por su simplicidad y su independencia del clasificador. Dentro del estado del arte, se encuentran métodos donde se cambia el proceso de aprendizaje o la función de costo para construir un clasificador robusto al desbalance [30–35]. Entre los anteriores se encuentran el *weighted lagrangian twin support vector machine*–(WLTSVM) [30], el cual es una variación de la máquina de vectores de soporte gemelas (TWSVM) [36], que tuvo resultados muy buenos en cuanto al tiempo de entrenamiento.
En este trabajo, proponemos un clasificador basado en máquinas de vectores de soporte gemelas que incluye una función kernel para mejorar la formulación del mismo para clasificación de datos desbalanceados. Nuestra propuesta es denominada *enhanced* TWSVM–(ETWSVM), la cual permite la representación de los datos de entrada en un espacio de alta dimensionalidad, posiblemente infinita dimensión, durante la optimización del clasificador. También, usamos un algoritmo basado en *centered kernel alignment*–(CKA) [37, 38] con el objetivo de aprender la función kernel de mapeo del ETWSVM para contrarrestar los problemas inherentes al desbalance y mejorar la separabilidad de los datos. Además, extendemos nuestro clasificador biclase a multiclase utilizando las estrategias One-versus-Rest (OvR) y One-versus-One (OvO) [39]. Como puntos de referencia para la clasificación binaria, probamos la SVM estándar [40], la SVM con variables slack regularizadas o WSVM [41], un SMOTE-SVM (SVM_{SMOTE}) [42], el mejorado TWSVM–(TBSVM) [12], y las máquinas de vectores de soporte gemelas con Lagrangiano ponderado (WLTSVM) [30]. Para la clasificación multi-clase, probamos SVM y el TBSVM con los esquemas OvR y OvO. Los resultados obtenidos sobre las bases de datos referentes del estado del arte muestran que nuestra propuesta supera los métodos de referencia en cuanto a las medidas de rendimiento de precisión, media geométrica y F-measure (F_1-score). También, estudiamos la influencia de los parámetros libre del ETWSVM considerando el desempeño del clasificador para diferentes tasas de desbalance y escenarios de traslape, sugerimos una variante de ETWSVM que obtiene un notable equilibrio entre desempeño y tiempo de entrenamiento, sin la necesidad de conocimiento a priori sofisticado con respecto a la sintonización del algoritmo.

El resto de este documento esta organizado como sigue: La parte I trata los temas preliminares, algoritmos de aprendizaje de gran importancia para nuestra propuesta. La parte II contiene lo relacionado a la teoría de nuestra propuesta y sus variaciones. Por último, la parte III contiene los resultados obtenidos con su respectiva discusión, además de las conclusiones y trabajos futuros.

6. Objetivos

6.1. Objetivos General

Desarrollar una metodología de clasificación basada en máquinas de vectores de soporte gemelas que considere el desbalance entre clases y que contemple la selección automática de parámetros libres en aras de favorecer la identificación de patrones relevantes con base en criterios de sensibilidad y especificidad.

6.2. Objetivos Específicos

- Proponer un método de clasificación basado en máquinas de vectores de soporte gemelas (TWSVM) que contemple mapeos no lineales con kernel reproductivo dentro de un esquema de optimización cuadrático, con el fin de favorecer la generalización de las representaciones utilizando distintos kernels según la aplicación de interés.
- Desarrollar una estrategia de sintonización de los parámetros libres del método de clasificación propuesto, que incluya la regularización automática del problema de aprendizaje y la escogencia de la función kernel a partir de los datos disponibles.
- Desarrollar una metodología de remuestreo para codificar estructuras de datos relevantes y favorecer el rendimiento en tareas de clasificación con desbalance, a partir del método de clasificación propuesto basado en TWSVM con selección automática de parámetros.

Parte I.

Preliminares

7. Preliminares Matemáticos

En este capitulo, proporcionamos una breve introducción a los conceptos de la teoría de funciones kernels para representación de muestras de entrada en el contexto de aprendizaje de máquina. Primero, se plantean las definiciones formales y necesarias, además de las condiciones suficientes para que una función sea representada por un *kernel reproductivo*. El contenido de este capitulo esta basado en artículos y libros de Parzen [43], Aronszajn [44], el Scholkopf y Smola [11].

7.1. Espacios de Hilbert con kernel reproductivo

Sea $\mathcal{X}$ un conjunto y $\mathcal{F}$ un espacio vectorial de funciones en $\mathcal{X}$ para el campo $\mathbb{F}$; en particular, $\mathbb{F}=\mathbb{R}$. Entonces, existe un espacio de Hilbert con kernel reproductivo(*reproductive kernel Hilbert space*–(RKHS)) $\mathcal{H}$ en $\mathcal{X}$ sobre $\mathbb{R}$, si:

- $\mathcal{H}$ es un sub-espacio vectorial de $\mathcal{F}$.
- $\mathcal{H}$ esta dotado de un producto punto, $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, y es completo dentro de la metrica inducida por este.
- Para cada $\boldsymbol{x}\in\mathcal{X}$ y $f\in\mathcal{H}$, la evolución lineal funcional $F_{\boldsymbol{x}} : \mathcal{H} \to \mathbb{R}$, se define como $F_{\boldsymbol{x}}(f)=f(\boldsymbol{x})$, es acotada

Desde el teorema de Riez [45], es de conocimiento que para cualquier función acotada H en un espacio de Hilbert $\mathcal{H}$, existe un único vector $\boldsymbol{h}\in\mathcal{H}$ tal que: $H(\boldsymbol{f})=\langle\boldsymbol{h},\boldsymbol{f}\rangle_{\mathcal{H}}\ \forall\boldsymbol{f}\in\mathcal{H}$. En consecuencia, para cada evaluación funcional $F_{\boldsymbol{x}}$ existe un correspondientes vector $\kappa_{\boldsymbol{x}}\in\mathcal{H}$. La función bivariable se define como

$$\kappa\big(\boldsymbol{x},\boldsymbol{x}'\big) = \kappa_{\boldsymbol{x}}(\boldsymbol{x}'), \tag{7.1}$$

esta función es llamada *kernel reproductivo* para $\mathcal{H}$, con $\boldsymbol{x}'\in\mathcal{X}$. Se puede verificar que:

$$\kappa\big(\boldsymbol{x},\boldsymbol{x}'\big) = \langle\kappa_{\boldsymbol{x}}, \kappa_{\boldsymbol{x}'}\rangle_{\mathcal{H}}\,, \tag{7.2}$$

y $\|F_{\boldsymbol{x}}\|^2_{\mathcal{H}}=\|\kappa_{\boldsymbol{x}}\|_{\mathcal{H}}=\langle\kappa_{\boldsymbol{x}}, \kappa_{\boldsymbol{x}'}\rangle_{\mathcal{H}}=\kappa(\boldsymbol{x},\boldsymbol{x}')$, donde $\|\cdot\|$ es el operador norma.

Sea $\mathcal{H}$ un RKHS en el conjunto $\mathcal{X}$ con kernel $\kappa(\cdot,\cdot)$. El tramo lineal de $\{\kappa(\boldsymbol{x},\cdot) : \boldsymbol{x}\in\mathcal{X}\}$ es denso en $\mathcal{H}$. Este resultado del hecho de que cualquier función f ortogonal para el tramo de $\{\kappa(\boldsymbol{x},\cdot) : \boldsymbol{x}\in\mathcal{X}\}$ debe satisfacer $\langle f, \kappa_{\boldsymbol{x}}\rangle_{\mathcal{H}}$, y así $f(\boldsymbol{x})=0$.

Lema 7.1. *Sea* $\{f_n\} \subset \mathcal{H}$*, con* $n\in\mathbb{N}$ *como un indice contable. Si* $\lim_{n\to+\infty}\|f_n - f\|_{\mathcal{H}}=0$*, entonces* $f(\boldsymbol{x})=\lim_{n\to+\infty} f_n(\boldsymbol{x})$ *para todo* $\boldsymbol{x}\in\mathcal{X}$.

Demostración. Esto es una simple consecuencia de la propiedad reproductiva y la desigualdad de Cauchy-Shwarz:

$$|f_n(\boldsymbol{x}) - f(\boldsymbol{x})| = |\langle f_n - f, \kappa_{\boldsymbol{x}}\rangle_{\mathcal{H}}| \leqslant \|f_n - f\|_{\mathcal{H}}\|\kappa_{\boldsymbol{x}}\|_{\mathcal{H}} \to 0$$

□

Proposición 7.2. *Sean $\mathcal{H}_1$ y $\mathcal{H}_2$ RKHS en $\mathcal{X}$ con funciones kernels κ_1 y κ_2, respectivamente. Si $\kappa_1(\boldsymbol{x},\boldsymbol{x}')=\kappa_2(\boldsymbol{x},\boldsymbol{x}')$ para todo $\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}$, entonces $\mathcal{H}_1=\mathcal{H}_2$ y $\|f\|_{\mathcal{H}_1}=\|f\|_{\mathcal{H}_2}$ $\forall f$.*

Demostración. Podemos tomar $\kappa(\boldsymbol{x},\boldsymbol{x}')=\kappa_1(\boldsymbol{x},\boldsymbol{x}')=\kappa_2(\boldsymbol{x},\boldsymbol{x}')$ y así el $M_l=\text{span}\{\kappa_{\boldsymbol{x}}\in M_l : \boldsymbol{x}\in\mathcal{X}\}$ es denso en $\mathcal{H}_l$, y para cualquier $f(\boldsymbol{x})=\sum_n \alpha_n\kappa_{\boldsymbol{x}_n}$ no existe consideración sobre si f pertenece a M_1 o a M_2. Note que $\|f\|^2_{\mathcal{H}_1}=\sum_{nn'}\alpha_n\alpha_{n'}\kappa(\boldsymbol{x},\boldsymbol{x}')=\|f\|^2_{\mathcal{H}_2}$ y así $\|f\|^2_{\mathcal{H}_1}=\|f\|^2_{\mathcal{H}_2}$ para todo $f\in M_1=M_2$. Si $f\in\mathcal{H}_1$, entonces existe una secuencia de funciones $\{f_n\}\subset M_1$ que converge a f en norma. Como, $\{f_n\}$ es Cauchy en M_1 es también Cauchy en M_2, así por lo completo de $\mathcal{H}_2$ existe $g\in\mathcal{H}_2$ tal que $f_n\to g$. Entonces, por el lema 7.1 tenemos que $f(\boldsymbol{x})=\lim_{n\to+\infty} f_n(\boldsymbol{x})=g(\boldsymbol{x})$ para todo $\boldsymbol{x}\in\mathcal{X}$, cada $f\in\mathcal{H}_1$ esta también en $\mathcal{H}_2$ y viceversa, y por ende $\mathcal{H}_1=\mathcal{H}_2$. Finalmente, podemos extender que $\|f\|_{\mathcal{H}_1}=\|f\|_{\mathcal{H}_2}$. □

Por lo tanto, dos RKHS diferentes no tienen el mismo kernel reproductivo. El siguiente teorema muestra un camino alterno para expresar el kernel reproductivo de un RKHS $\mathcal{H}$.

Teorema 7.3. *Sea $\mathcal{H}$ un RKHS con kernel reproductivo κ. Si $\{e_\lambda(\boldsymbol{x}) : \lambda\in\Lambda\}$ es una base ortonormal de $\mathcal{H}$, entonces:*

$$\kappa(\boldsymbol{x},\boldsymbol{x}')=\sum_{\lambda\in\Lambda} e_\lambda(\boldsymbol{x})e_\lambda(\boldsymbol{x}'), \tag{7.3}$$

donde las series convergen de forma puntual.

Demostración. Para un conjunto fijo $\{\boldsymbol{x}_n\}\subseteq\mathcal{X}$, tenemos:

$$\sum_{n,n'=1}^{N}\alpha_n\alpha_{n'}\kappa(\boldsymbol{x}_n,\boldsymbol{x}_{n'}) = \left\langle \sum_{n}^{N}\alpha_n\kappa_{\boldsymbol{x}_n}, \sum_{n'}^{N}\alpha_{n'}\kappa_{\boldsymbol{x}_{n'}}\right\rangle_{\mathcal{H}} = \left\|\sum_{n}^{N}\alpha_n\kappa_{\boldsymbol{x}_n}\right\|_{\mathcal{H}} \geqslant 0$$

□

Adicional a esto, el teorema de Moore es introducido, el cual es contrario al anterior resultado y nos proporciona la característica de una función definida positiva como condición suficiente para que la función sea un kernel reproductivo de algún RKHS.

Teorema 7.4. *Sea $\mathcal{X}$ un conjunto y $\kappa : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ una función definida positiva. Entonces, existe un RKHS $\mathcal{H}$ de funciones en $\mathcal{X}$, tal que, κ es un kernel reproductivo de $\mathcal{H}$.*

Demostración. Considere las funciones $\kappa_{\boldsymbol{x}}(\boldsymbol{x}')=\kappa(\boldsymbol{x},\boldsymbol{x}')$ y el espacio W extendido por el conjunto $\{\kappa_{\boldsymbol{x}} : \boldsymbol{x}\in\mathcal{X}\}$. El siguiente mapeo bilineal $B : W\times W\to\mathbb{R}$:

$$B\left(\sum_n \alpha_n\kappa_{\boldsymbol{x}_n}, \sum_{n'}\beta_{n'}\kappa_{\boldsymbol{x}_{n'}}\right) = \sum_{n,n'}\alpha_n\beta_{n'}\kappa(\boldsymbol{x}_n,\boldsymbol{x}_{n'}),$$

donde $\alpha_n,\beta_{n'}\in\mathbb{R}$, esta bien definido en W. Para respaldar lo dicho anteriormente, notese que si $f(\boldsymbol{x})=\sum_n\alpha_n\kappa_{\boldsymbol{x}_n}(\boldsymbol{x})$ es cero para todo $\boldsymbol{x}\in\mathcal{X}$, entonces por definición $B(f,\kappa_{\boldsymbol{x}})=0$ $\forall\boldsymbol{x}$. A la inversa, si $B(f,w)=0$ para todo $w\in W$, entonces tomando $w=\kappa_{\boldsymbol{x}}$ se puede ver que $f(\boldsymbol{x})=0$. Entonces, B esta bien definida.

Como κ es definida positiva $B(f,f) \geqslant 0$ y vemos que $B(f,f)=0$ si y sólo si $B(w,f)=0$ para todo $w \in W$, por lo tanto $f(\boldsymbol{x})=0$ para todo $\boldsymbol{x} \in \mathcal{X}$. Ahora, hemos visto que W es un sub-espacio pre-Hilbert con producto interno B. Dejemos que $\mathcal{H}$ denote la terminación de W, necesitamos ver que cada elemento de $\mathcal{H}$ es función sobre $\mathcal{X}$. Sea $h \in \mathcal{H}$ el punto limite de una secuencia de Cauchy $\{f_n\} \subseteq W$. Por la desigualdad de Cauchy-Shwarz:

$$|f_n(\boldsymbol{x}) - f_{n'}(\boldsymbol{x})| = |B(f_n - f_{n'}, \kappa_{\boldsymbol{x}})| \leqslant \|f_n - f_{n'}\| \kappa(\boldsymbol{x},\boldsymbol{x}) \,.$$

Por lo tanto, el limite puntual $h(\boldsymbol{x}) = \lim_{n \to +\infty} f_n(\boldsymbol{x})$ esta bien definida. Concluyendo, sea $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ el producto interior sobre $\mathcal{H}$. Entonces, tenemos $\langle h, \kappa_{\boldsymbol{x}} \rangle_{\mathcal{H}} = \lim_{n \to +\infty} \langle f_n, \kappa_{\boldsymbol{x}} \rangle_{\mathcal{H}} = B(f_n, \kappa_{\boldsymbol{x}}) = h(\boldsymbol{x})$. Así $\mathcal{H}$ es un RKHS con kernel reproductivo κ. □

Combinando la proposición 7.2 con el teorema de Moore (teorema 7.4) vemos la correspondencia entre RKHS sobre el conjunto $\mathcal{X}$ y funciones definidas positivas sobre este conjunto.

7.2. Espacios de Hilbert con kernel reproducivo en aprendizaje de máquina

Es de conocimiento general que el estudio de funciones kernels definidas positivos es un tema de interés para la comunidad del aprendizaje de máquina como una generalización de un buen cuerpo teórico que ha sido desarrollado para modelos lineales. Una función kernel definida positiva κ, es un camino implícito para representar las muestras de un espacio de entrada $\mathcal{X}$. Debido a que hay una correspondencia entre κ y los RKHS de funciones en $\mathcal{H}$, la función kernel puede ser entendida como un camino indirecto para el calculo del producto interno entre elementos de un espacio de Hilbert, los cuales son resultado de elementos mapeados de $\mathcal{X}$ a $\mathcal{H}$. Por ende, existe una función de mapeo $\varphi : \mathcal{X} \to \mathcal{H}$ tal que:

$$\kappa(\boldsymbol{x},\boldsymbol{x}') = \langle \varphi(\boldsymbol{x}) \,, \varphi(\boldsymbol{x}') \rangle_{\mathcal{H}} \,. \tag{7.4}$$

Con respecto a esto, el espacio $\mathcal{H}$ puede ser visto como un espacio de características y φ es conocida la función de mapeo al espacio de características. Como consecuencia, mediante la realización de operaciones lineales en $\mathcal{H}$ es posible realizar manipulaciones lineales en el espacio de entrada $\mathcal{X}$; no obstante, sin necesidad de realizar cualquier calculo explicito en $\mathcal{H}$ (ver figura 7.1).

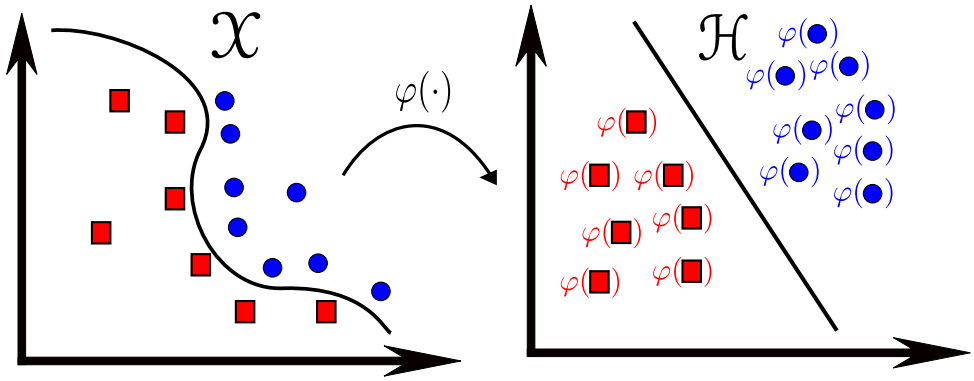


Figura 7.1.: Mapeo basado en función kernel.

Entonces, una importante propiedad asociada con el uso de funciones kernels, definidas positivas, en aprendizaje de máquina, es conocido como el *teorema de representación* [11, 46].

Teorema 7.5. *Sea $\Omega : [0, +\infty) \to \mathbb{R}$ una función monótonamente creciente de forma estricta, $\mathcal{X}$ un conjunto, y $\epsilon : (\mathcal{X} \times \mathbb{R}^2)^N \to \mathbb{R} \cup \infty$ una función de perdida arbitraria. Entonces, cada minimizador $f \in \mathcal{H}$ del riesgo regularizado funcional:*

$$\epsilon((\boldsymbol{x}_1, t_1, f(\boldsymbol{x}_1)), \ldots, (\boldsymbol{x}_N, t_N, f(\boldsymbol{x}_N))) + \Omega\Big(\|f\|^2_{\mathcal{H}}\Big) \tag{7.5}$$

admite una representación de la forma

$$f(\boldsymbol{x}) = \sum_{n=1}^{N} \alpha_n \kappa(\boldsymbol{x}_n, \boldsymbol{x}) \,, \tag{7.6}$$

donde cada $t_n \in \mathbb{R}$ es un salida asociada con la entrada $\boldsymbol{x}_n \in \mathcal{X}$.

Demostración. Sea $S = \text{span}\{\kappa(\boldsymbol{x}_n, \cdot) : \boldsymbol{x}_n \in \mathcal{X}, n \in \{1, \ldots, N\}\}$ el subespacio extendido $\mathcal{H}$ para las N muestras de entrenamiento. Considere la solución $f \in \mathcal{H}$, esta solución puede ser escrita como: $f = f_S + f_{S^\perp}$, donde $f_S \in S$, $f_{S^\perp} \in S^\perp$, y $\perp$ indica el símbolo ortogonal. Consecuentemente, $f(\boldsymbol{x}_n) = f_S(\boldsymbol{x}_n) + f_{S^\perp}(\boldsymbol{x}_n) = f_S(\boldsymbol{x}_n) + 0$. Ahora, para el segundo termino del riesgo regularizado funcional:

$$\Omega\Big(\|f\|^2_{\mathcal{H}}\Big) = \Omega\Big(\|f_S\|^2_{\mathcal{H}} + \|f_{S^\perp}\|^2_{\mathcal{H}}\Big),$$

como Ω es estrictamente monótona creciente, es posible ver que el mínimo será registrado por $\|f_{S^\perp}\|_{\mathcal{H}} = 0$, lo cual implica que $f_{S^\perp} = 0$.

□

Con esto en mente, es posible concluir que el teorema de representación básicamente declara que la solución de la funcion de riesgo regularizado puede ser expresada en términos de las llamadas muestras de entrenamiento $\{\boldsymbol{x}_n, t_n\}_{n=1}^N$. Por lo tanto, nos permite abordar con problemas que a simple vista parecen ser de dimensión infinita. No obstante, la regularización previene de múltiples mínimos locales, tal propiedad requiere algunas condiciones extras, normalmente, convexidad.

8. Multiplicadores de Lagrange

Los multiplicadores de Lagrange son usados para encontrar puntos estacionarios de una función sujeta a varias variables para una o más restricciones.Consideremos el problema de hallar el máximo de la función $f(x_1, x_2)$ sujeto a una restricción relativa a x_1 y x_2, el cual vamos ha escribir de la forma:

$$g(x_1, x_2) = 0. \tag{8.1}$$

Un enfoque sería resolver la restricción de la ecuación (8.1) y entonces expresar x_2 en función de x_1 de la forma $x_2{=}h(x_1)$. Este puede ser entonces sustituido dentro de $f(x_1, x_2)$ para obtener una función dependiente unicamente de x_1, tomando la forma $f(x_1, h(x_1))$. Luego, el máximo con respecto a x_1 puede ser encontrado por el camino usual de la diferenciación, para obtener el valor estacionario $x_1^\star$, y con este valor calcular $x_2^\star{=}h(x_1^\star)$.

Uno de los problema del anterior enfoque es la dificultad que puede presentar encontrar la solución analítica de la ecuación de la restricción que permite a x_2 ser expresado como una función explicita de x_1. También, este enfoque trata a x_1 y x_2 de forma diferente y así estropea la simetría natural entre estas variables.

Un más elegante, y con frecuencia más simple, enfoque es basado en la introducción de un parámetro λ llamado un *multiplicador de Lagrange*. Motivaremos esta técnica desde una perspectiva geométrica. Consideremos un vector $\boldsymbol{x}{\in}\mathbb{R}^D$ con la ecuación de restricción $g(\boldsymbol{x}){=}0$, la cual representa una superficie de $(D-1)$–dimensiones en el espacio de $\boldsymbol{x}$ como es indicado en la figura 8.1(a).

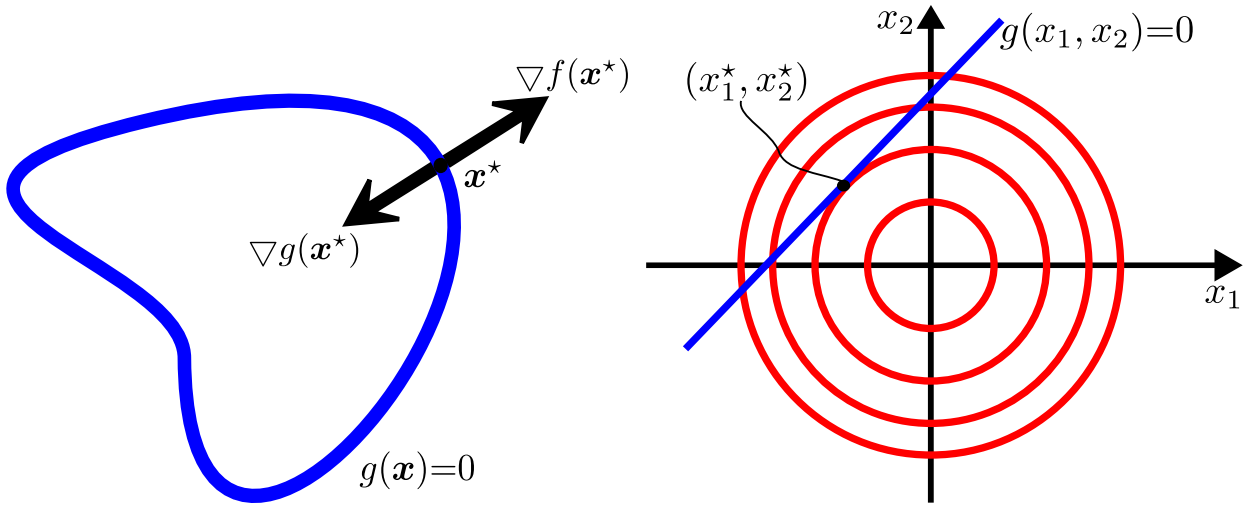
Primero debemos notar que cualquier punto, sobre la superficie de la restricción, evaluado en el gradiente $\bigtriangledown g(\boldsymbol{x})$ será ortogonal a la superficie generada por $g(\boldsymbol{x}){=}0$. Para observar esto, considere un punto $\boldsymbol{x}$ que se encuentra sobre la superficie de restricción, y considere un punto vecino $\boldsymbol{x}+\boldsymbol{\epsilon}$ que también se encuentra sobre la superficie. Si realizamos una expansión de Taylor alrededor de $\boldsymbol{x}$, tenemos

$$g(\boldsymbol{x}+\boldsymbol{\epsilon}) \simeq g(\boldsymbol{x}) + \boldsymbol{\epsilon}^\top \bigtriangledown g(\boldsymbol{x}). \tag{8.2}$$

Debido a que ambos $\boldsymbol{x}$ y $\boldsymbol{x}+\boldsymbol{\epsilon}$ se encuentran sobre la superficie de restricción, tenemos que $g(\boldsymbol{x}){=}g(\boldsymbol{x}+\boldsymbol{\epsilon}){=}0$, y por lo tanto $\boldsymbol{\epsilon}^\top\bigtriangledown g(\boldsymbol{x}) \simeq 0$. Si el limite $\|\boldsymbol{\epsilon}\|_2 \to 0$, entonces $\boldsymbol{\epsilon}^\top\bigtriangledown g(\boldsymbol{x}){=}0$, y porque $\boldsymbol{\epsilon}$ es entonces paralelo a la superficie de restricción, vemos que el vector $\bigtriangledown g$ es normal a esta superficie.

Luego buscamos un punto $\boldsymbol{x}^\star$ sobre la superficie de restricción tal que $f(\boldsymbol{x}^\star)$ sea máximo. Tal punto debe tener la propiedad de que el vector $\bigtriangledown f(\boldsymbol{x}^\star)$ sea también ortogonal a la superficie de restricción, $g(\boldsymbol{x}){=}0$, como se ilustra en la figura 8.1(a), dado que en cualquier otro caso podríamos incrementar el valor de $f(\boldsymbol{x}^\star)$ moviendo una corta distancia sobre la superficie de restricción. De este modo, $\bigtriangledown f(\boldsymbol{x}^\star)$ y $\bigtriangledown g(\boldsymbol{x}^\star)$ son vectores anti-paralelos, y entonces debe de existir un parámetro λ tal que

$$\bigtriangledown f(\boldsymbol{x}^\star) + \lambda \bigtriangledown g(\boldsymbol{x}^\star) = 0, \tag{8.3}$$



(a) Representación Geometrica (b) Ejemplo

Figura 8.1.: Figura geométrica de la técnica de multiplicadores de Lagrange en el cual buscamos maximizar una función $f(\boldsymbol{x})$, sujeta a la restricción $g(\boldsymbol{x})=0$. Si $\boldsymbol{x}$ es de dimensión D, la restricción corresponde a un subespacio de dimensión $D-1$, indicada por la curva azul. El problema puede resolverse por optimización de la función Lagrangiana $\mathcal{L}(\boldsymbol{x}, \lambda)=f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$. Derecha: Un ejemplo sencillo del uso de los multiplicadores de Lagrange en el cual tiene el objetivo de maximizar $f(x_1, x_2)$ sujeto a la restricción $g(x_1, x_2)=0$. Los círculos rojos muestran las curvas de nivel de la función $f(x_1, x_2)$, y la linea azul corresponde a la superficie de restricción $g(x_1, x_2)=0$.

donde $\lambda \neq 0$ es un *multiplicador de Lagrange*. Notemos que λ puede tener cualquier signo.

En este punto, es conveniente introducir la función del *Lagrangiano* definida por:

$$\mathcal{L}(\boldsymbol{x}, \lambda) \equiv f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}). \tag{8.4}$$

La condición de la restricción estacionaria (ecuación (8.3)) es obtenida por medio de la igualdades $\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \lambda)=0$ y $\partial \mathcal{L}(\boldsymbol{x}, \lambda)/\partial \lambda=0$.

De este modo, para encontrar el máximo de la función $f(\boldsymbol{x})$ sujeta a la restricción $g(\boldsymbol{x})=0$, definimos la función del Lagrangiano dada en la ecuación (8.4) y entonces encontramos el punto estacionario de $\mathcal{L}(\boldsymbol{x}, \lambda)$ con respecto a $\boldsymbol{x}$ y λ. Para un vector $\boldsymbol{x} \in \mathbb{R}^D$, este otorga $D+1$ ecuaciones las cuales determinan el punto estacionario $\boldsymbol{x}^*$ y el valor del multiplicador de Lagrange λ.

Como un simple ejemplo, supongamos que deseamos encontrar el punto estacionario de la función $f(x_1, x_2)$ $=1 - x_1^2 - x_2^2$ sujeto a la restricción $g(x_1, x_2)=x_1 + x_2 - 1=0$, como se ilustra en la figura 8.1(b). La correspondiente función del Lagrangiano esta dada por:

$$\mathcal{L}(\boldsymbol{x}, \lambda) = 1 - x_1^2 - x_2^2 + \lambda(x_1 + x_2 - 1). \tag{8.5}$$

Las condiciones para que este Lagrangiano sea estacionario con respecto a x_1, x_2, y λ están dadas por las siguientes ecuaciones:

$$-2x_1 + \lambda = 0 \tag{8.6}$$
$$-2x_2 + \lambda = 0 \tag{8.7}$$
$$x_1 + x_2 - 1 = 0. \tag{8.8}$$

Solucionando las ecuaciones obtenemos el punto estacionario como $[x_1^\star, x_2^\star]$=$[\frac{1}{2}, \frac{1}{2}]$,y el correspondiente valor para el multiplicador de Lagrange es λ=1.

Hasta el momento, hemos considerado un problema de minimización sujeta a una *restricción de igualdad* de la forma $g(\boldsymbol{x})$=0. Ahora consideraremos el problema de maximizar $f(\boldsymbol{x})$ sujeto a una *restricción de desigualdad* de la forma $g(\boldsymbol{x}) \geqslant 0$, como se ilustra en la figura 8.2.

Ahora existen dos tipos de soluciones posibles, acorde a la restricción estacionaria, que son los puntos que se encuentran dentro de la región $g(\boldsymbol{x}) > 0$, en tal caso la restricción es *inactiva*, o los que se encuentran sobre la restricción $g(\boldsymbol{x})$=0, en este caso se dice que la restricción es *activa*. En el primer caso, la función $g(\boldsymbol{x})$ no juega ningún papel y la condición estacionaria es simplemente $\bigtriangledown f(\boldsymbol{x})$=0. Este de nuevo corresponde al punto estacionario de la función de Lagrange definida en la ecuación (8.4), pero con λ=0. El otro caso, donde la solución está sobre el límite, es análogo a la restricción de igualdad discutido previamente y corresponde al punto estacionario de la función de Lagrange con $\lambda \neq 0$. Sin embargo, el signo del multiplicador de Lagrange es crucial, debido a que la función $f(\boldsymbol{x})$ sólo puede ser máxima si su gradiente es orientado lejos de la región $g(\boldsymbol{x}) > 0$, como se ilustra en la figura 8.2. Por lo tanto, tenemos $\bigtriangledown f(\boldsymbol{x}) = -\lambda \bigtriangledown g(\boldsymbol{x})$ para algunos valores de $\lambda > 0$.

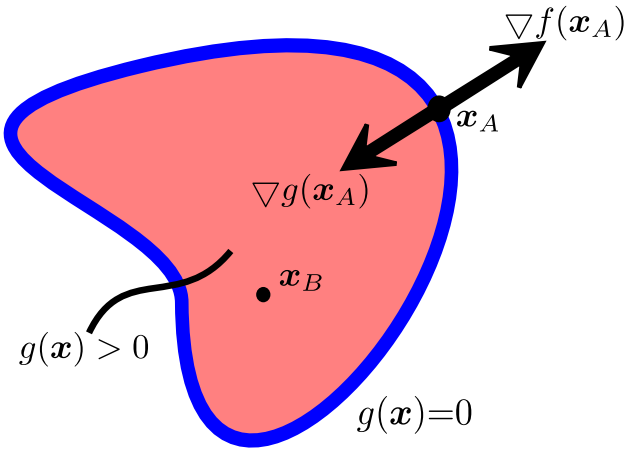


Figura 8.2.: Representación de el problema de minimizar $f(\boldsymbol{x})$ sujeto a la restricción de desigualdad $g(\boldsymbol{x}) \geqslant 0$.

Para cualquiera de estos dos casos, el producto $\lambda g(\boldsymbol{x})$=0. Así, la solución para el problema de minimización de $f(\boldsymbol{x})$ sujeta a la restricción $g(\boldsymbol{x}) \geqslant 0$ es obtenida por la optimización de la función de Lagrange (ecuación (8.4)) con respecto a $\boldsymbol{x}$ y λ sujeta a las condiciones

$$g(\boldsymbol{x}) \geqslant 0 \tag{8.9}$$
$$\lambda \geqslant 0 \tag{8.10}$$
$$\lambda g(\boldsymbol{x}) = 0. \tag{8.11}$$

Estas condiciones son conocidas como las *condiciones de Karush-Kuhn-Tucker*–(KKT) [47, 48].

Notemos que si deseamos minimizar la función $f(\boldsymbol{x})$ sujeto a la restricción de desigualdad $g(\boldsymbol{x}) \geqslant 0$, sólo es necesario minimizar la función del Lagrangiano $\mathcal{L}(\boldsymbol{x}, \lambda)=f(\boldsymbol{x}) - \lambda g(\boldsymbol{x})$ con respecto a $\boldsymbol{x}$, y de nuevo sujeto a que $\lambda \geqslant 0$.

Finalmente, para extender la técnica de multiplicadores de Lagrange al caso de múltiples restricciones de igualdad y desigualdad, suponemos que deseamos maximizar la función $f(\boldsymbol{x})$ sujeto a $g_j(\boldsymbol{x})=0$ para $j\in\{1, 2, \ldots, J\}$, y $h_k(\boldsymbol{x}) \geqslant 0$ para $k\in\{1, 2, \ldots, K\}$. Introducimos entonces los multiplicadores de Lagrange $\{\lambda_j\}_{j=1}^{J}$ y $\{\mu_k\}_{k=1}^{K}$, y entonces optimizamos la función del Lagrangiano dada por

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{j=1}^{J} \lambda_j g_j(\boldsymbol{x}) + \sum_{k=1}^{K} \mu_k h_k(\boldsymbol{x}), \tag{8.12}$$

sujeto a que $\mu_k \geqslant 0$ y $\mu_k h_k(\boldsymbol{x})=0$ para $k\in\{1, \ldots, K\}$.

9. Preliminares Clasificación

9.1. Aprendizaje supervisado, clasificación desbalanceadas

Para nosotros es fácil reconocer un rostro, entender palabras habladas, leer caracteres realizados a mano alzada y decidir sobre una naranja o una mandarina por su olor; esto, ha sido indispensable para nuestra supervivencia en los pasados diez mil años. El reconocimiento de patrones se define por [49] como: *"el acto de tomar los datos sin procesar y realizar una acción basada en categoría del patrón"*. Debido, al avance tecnológico hemos desarrollado sistemas neuronales y cognitivos sofisticados para realizar esta tarea a gran escala.

Dicho lo anterior, dentro del reconocimiento de patrones se encuentra la tarea de clasificación, la cual consiste en asignar una categoría o clase a un elemento, o muestra, a partir de características de los datos. Un ejemplo de clasificación binara, es el problema de un granjero para separar por medio de una cerca a sus caballos de sus vacas con el fin de tenerlos en dos zonas de pastoreo diferentes. El inconveniente es determinar cuál es la mejor dirección de la cerca para esta tarea. La figura 9.1 ejemplifica la situación mencionada; podemos observar que existen varias opciones para separar ambas especies, una de ellas es por medio de la cerca azul y otra por la roja, esto nos deja la pregunta ¿cuál es la mejor opción?. Y así determinar a partir de la zona si un animal es de una especie o de la otra.

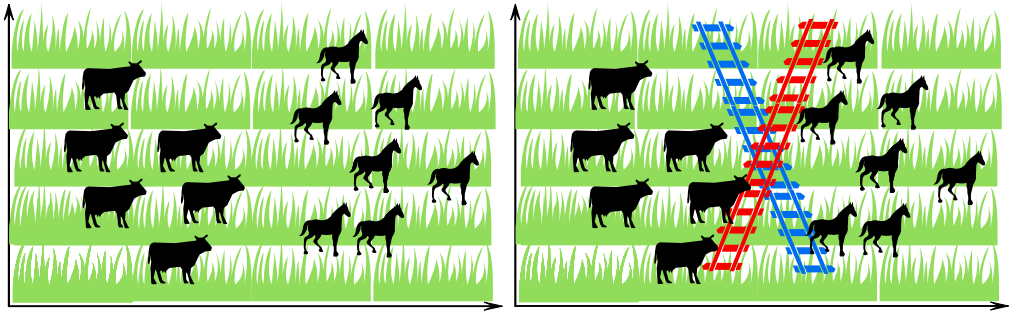


Figura 9.1.: Ejemplo de clasificación

Para realizar esta tarea desde un punto de vista estadístico y/o probabilístico se hace uso del aprendizaje de máquina, el cual definimos como un conjunto de métodos que pueden detectar patrones en datos de forma automática, y que luego usa estos patrones detectados para predecir un dato futuro (nuevo), o realizar otro tipo de toma de decisiones bajo incertidumbre; por ejemplo, decidir si un dato pertenece a alguna clase en especifico [50]. En este trabajo nos concentramos en las famosas Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés) propuestas por [51]. Este método de clasificación y regresión es famoso por su capacidad de generalización y su uso en diferentes tareas del mundo real como identificación de

partículas, categorización de texto, bio-informática, y aplicaciones financieras [52–54]. Este algoritmo de aprendizaje en su etapa de entrenamiento debe resolver un problema de optimización cuadrático (*quadratic programming problem*–(QPP)), además su complejidad computacional es de $\mathcal{O}(N^3)$, siendo N el número de datos a utilizar en esta etapa.

9.2. Máquinas de vectores de Soporte (SVM)

Sea $\{\boldsymbol{X}\in\mathbb{R}^{P\times N}, \boldsymbol{t}\in\{-1,+1\}^N\}$ una base de datos con N muestras de entrada $\boldsymbol{x}_n\in\mathbb{R}^P$ y salidas etiquetadas $t_n\in\{-1,+1\}$, con $n\in\{1,2,\ldots,N\}$, para clasificación binaria. La SVM busca encontrar un hiperplano de la forma:

$$f(\boldsymbol{x})=\boldsymbol{w}^\top\boldsymbol{x}+b=0, \tag{9.1}$$

donde $\boldsymbol{w}\in\mathbb{R}^P$ es un vector normal ponderado del hiperplano y $b\in\mathbb{R}$ es un termino de sesgo. Además, la SVM escoge la frontera de decisión (hiperplano) que maximizá la margen, la cual es definida como la mínima distancia perpendicular entre el hiperplano y cualquiera de las muestras. Pero, primero debemos definir la expresión para la distancia perpendicular de una muestra $\boldsymbol{x}$ al hiperplano. Podemos demostrar que el valor de $f(\boldsymbol{x})$ da una medida con signo de la distancia perpendicular r del punto $\boldsymbol{x}$ desde la superficie de decisión. Para observar esto, consideremos el escenario propuesto por la figura 9.2.

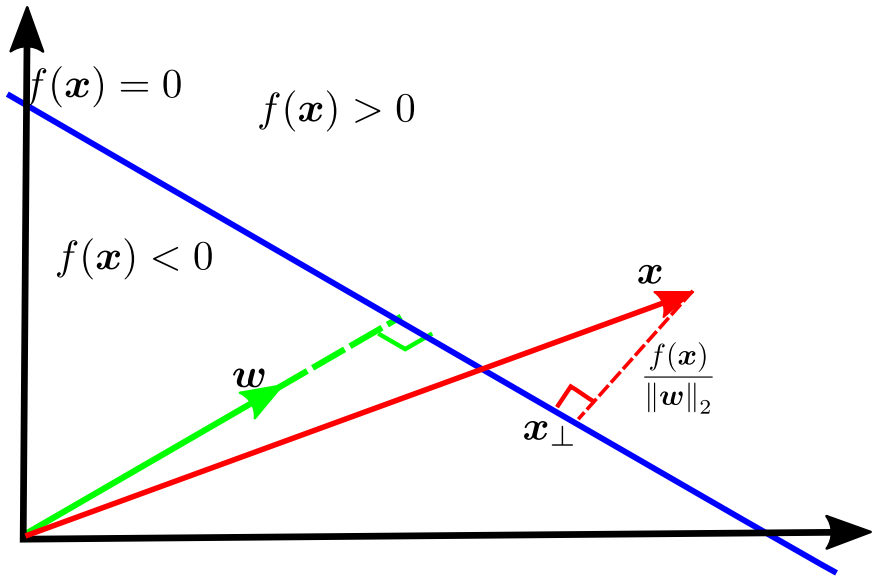


Figura 9.2.: Ilustración geométrica de la distancia ortogonal de una muestra $\boldsymbol{x}$, representada en color rojo y la superficie de decisión,vista en azul; su vector perpendicular $\boldsymbol{w}$ es ilustrado en verde.

También consideremos un punto arbitrario $\boldsymbol{x}$ y sea $\boldsymbol{x}_\perp$ su proyección ortogonal sobre el hiperplano $f(\boldsymbol{x})=0$, entonces podemos decir que

$$\boldsymbol{x}=\boldsymbol{x}_\perp+r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}, \tag{9.2}$$

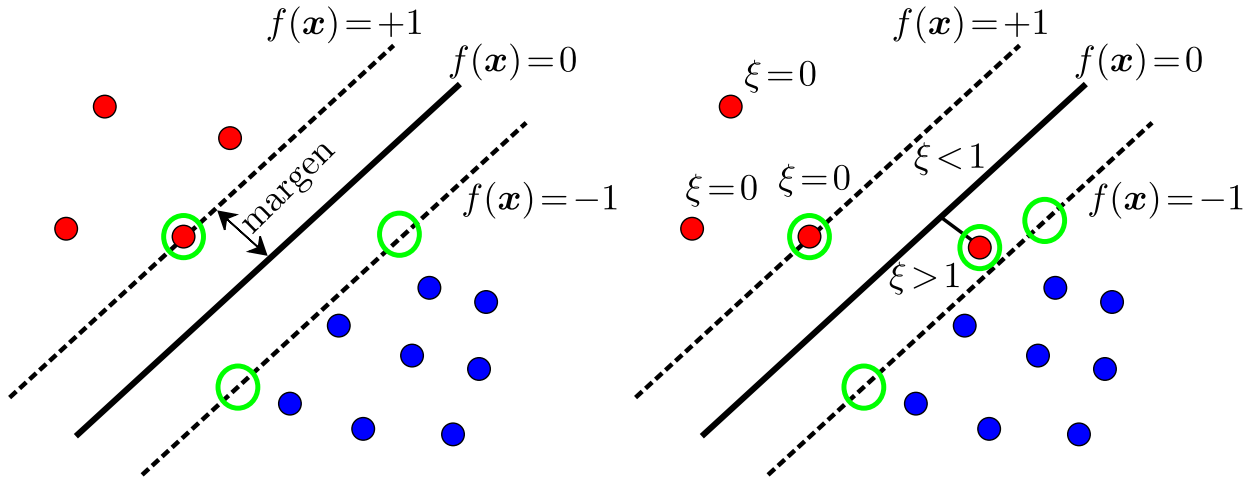
donde $\|\cdot\|_2$ es el operador de norma L_2.
Ahora multiplicando ambos lados por el término $\boldsymbol{w}^\top$ y luego sumamos b a ambos lados, y usando las expresión $f(\boldsymbol{x})=\boldsymbol{w}^\top\boldsymbol{x}+b$ y $f(\boldsymbol{x}_\perp)=\boldsymbol{w}^\top\boldsymbol{x}_\perp+b=0$, obtenemos que

$$r = \frac{f(\boldsymbol{x})}{\|\boldsymbol{w}\|_2}. \tag{9.3}$$

Ahora, esta expresión puede llegar a poseer signo negativo dependiendo de la ubicación del punto $\boldsymbol{x}$; por esto, hacemos uso del operador valor absoluto $|\cdot|$ y la expresión para calcular la distancia perpendicular de una muestra al hiperplano queda de la forma

$$\frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|_2} = \frac{t_n f(\boldsymbol{x})}{\|\boldsymbol{w}\|_2} = \frac{t_n\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)}{\|\boldsymbol{w}\|_2}. \tag{9.4}$$

Siguiendo con la formulación en la SVM, la figura 9.3(a) muestra un ejemplo gráfico de la mismas e ilustra la definición de margen. También, observamos muestras pertenecientes a dos clases, una clase representada con el color rojo y otra por el azul. Podemos ver en esta figura, que la margen la define la muestra roja más cercana, ortogonalmente, al hiperplano; adicionalmente, este punto de alguna manera definirá el hiperplano y por esto vendría hacer parte del subconjunto de muestras llamado, *vectores de soporte*–(SV, por sus siglas en inglés).



(a) Margen y SV

(b) Variable Slack

Figura 9.3.: La SVM busca maximizar la margen, la cual se define como la mínima distancia perpendicular del hiperplano a cualquiera de las muestras disponibles. (a) muestra un ejemplo gráfico sobre la definición del la *magen*. (b) es un ejemplo donde existe traslape entre las clases y se permite un hiperplano que clasifique muestras de una clase que no se encuentren en zona correspondiente, para ello se hace uso de las variables auxiliares ξ_n. En ambas figuras, las muestras encerradas en un círculo verde son consideradas vectores de soporte y son las que determinan el hiperplano.

Por otra parte, cuando las muestras de ambas clases se traslapan definir una frontera de decisión lineal es una tarea casi imposible. Por lo anterior, las SVM consideran variables auxiliares $\xi_n{=}|t_n - f(\boldsymbol{x}_n)|$, para $n{\in}\{1, 2, \ldots, N\}$. Las muestras correspondientes a $\xi_n{=}0$ estarían bien clasificadas y a una distancia mayor que la margen; por otra parte, las muestras donde $\xi_n{=}1$ indicaría que están sobre la frontera de decisión; y por último, las muestras con $\xi_n > 1$, quieren decir que se encuentran mal clasificadas por la frontera de decisión. Estas variables auxiliares dan a este modelo de una tolerancia ante la mala clasificación, de algunas muestras difíciles, como lo son por ejemplo las que se traslapan con la otra clase, y así generar una frontera de decisión más generalizante. Por este motivo, el conjunto de muestras que definen la margen con $\xi_n > 1$

son considerados vectores de soporte.

Asimismo, nos interesa que todos las muestras queden correctamente clasificadas, para que $t_n f(\boldsymbol{x}_n) > 0 \; \forall_{n=1,2,\ldots,N}$. La margen esta dada como la distancia perpendicular al punto más cercano $\boldsymbol{x}_n$ del conjunto de datos, y deseamos optimizar los parámetros $\boldsymbol{w}$ y b para maximizar esta distancia, Así, la solución para la máxima margen se encuentra resolviendo,

$$\boldsymbol{w}^*, b^* = \arg\max_{\boldsymbol{w},b} \left\{ \frac{1}{\|\boldsymbol{w}\|_2} \min_n \left[t_n \left(\boldsymbol{w}^\top \boldsymbol{x}_n + b \right) \right] \right\} \tag{9.5}$$

donde el factor $1/\|\boldsymbol{w}\|_2$ se factoriza del problema de optimización sobre n, dado que no depende de n. La solución directa de este problema de optimización es compleja; por ende, se cambia el problema de optimización de la margen a uno más sencillo de resolver. Para lo anterior, debemos de notar que si re-escalamos los parámetros del hiperplano, $\boldsymbol{w} \to \alpha \boldsymbol{w}$ y $b \to \alpha b$, entonces la distancia de cualquier muestra $\boldsymbol{x}_n$ a la frontera de decisión, $t_n f(\boldsymbol{x}_n)/\|\boldsymbol{w}\|_2$, no cambia y todas las muestras satisfarán la restricción

$$t_n \Big(\boldsymbol{w}^\top \boldsymbol{x}_n + b \Big) \geqslant 1 - \xi_n, \quad \forall n \in \{1, 2, \ldots, N\}. \tag{9.6}$$

Lo anterior, es conocido como representación canónica del hiperplano de decisión. En el caso de las muestras que mantienen la igualdad, se dice que las restricciones están activas, mientras que para las demás muestras se dice que están inactivas. Por la definición de la margen, siempre habrá al menos una muestra activa. El problema de optimización simplemente requerirá que maximicemos $\|\boldsymbol{w}\|_2^{-1}$, lo cual es equivalente a minimizar $\|\boldsymbol{w}\|_2^2$, entonces tenemos que resolver el siguiente problema de optimización:

$$\begin{aligned} \widetilde{\boldsymbol{w}}, \widetilde{b} = \arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C \sum_{n=1}^{N} \xi_n \\ \text{s.t. } & t_n \Big(\boldsymbol{w}^\top \boldsymbol{x}_n + b \Big) \geqslant 1 - \xi_n, \\ & \xi_n \geqslant 0 \end{aligned} \tag{9.7}$$

$\forall n \in \{1, 2, \ldots, N\}$; donde $C \in \mathbb{R}^+$ es el parámetro que regulariza la penalización de las variables auxiliares y la margen. Debido a que cualquier punto mal clasificado tiene $\xi_n > 1$, además es importante recalcar que al minimizar el termino $\|\boldsymbol{w}\|_2^2$ junto con la restricción expresada en la ecuación (9.6), se maximiza de una manera menos compleja la margen del hiperplano; por otra parte, el minimizar el termino $\sum_n \xi_n$ ayuda a mitigar los escenarios de traslape entre clases y así reducir en lo posible la mala clasificación de las muestras de entrenamiento. El lagrangiano (ver capítulo 8) del problema de optimización expresado en la ecuación (9.7) esta dado por:

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C \sum_{1=1}^{N} \xi_n - \sum_{1=1}^{N} \alpha_n (t_n f(\boldsymbol{x}_n) - 1 + \xi_n) - \sum_{1=1}^{N} \mu_n \xi_n, \tag{9.8}$$

donde $\{\alpha_n \geqslant 0\}$ y $\{\mu_n \geqslant 0\}$ son multiplicadores de Lagrange, y $\boldsymbol{\xi} \in \mathbb{R}^N$ es el vector de variables auxiliares con elementos ξ_n. Su correspondiente conjunto de condiciones *Karush-Kuhn-Tucker* (KKT) (ver capítulo 8) son expresadas por:

$$\alpha_n \geqslant 0 \tag{9.9}$$
$$t_n f(\boldsymbol{x}_n) - 1 + \xi_n \geqslant 0 \tag{9.10}$$
$$\alpha_n(t_n f(\boldsymbol{x}_n) - 1 + \xi_n) = 0 \tag{9.11}$$
$$\mu_n \geqslant 0 \tag{9.12}$$
$$\xi_n \geqslant 0 \tag{9.13}$$
$$\mu_n \xi_n = 0, \tag{9.14}$$

$\forall n \in \{1, 2, \ldots, N\}$.
Ahora optimizamos $\boldsymbol{w}$, b, y $\{\xi_n\}_{n=1}^N$ haciendo uso de la definición para $f(\boldsymbol{x})$ dada en la ecuación (9.1), con lo que obtenemos:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \mathbf{0} \Rightarrow \boldsymbol{w} = \sum_{n=1}^{N} \alpha_n t_n \boldsymbol{x}_n \tag{9.15}$$
$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n t_n = 0 \tag{9.16}$$
$$\frac{\partial \mathcal{L}}{\partial \xi_n} = 0 \Rightarrow \alpha_n = C - \mu_n. \tag{9.17}$$

Usando estos resultados con el fin de eliminar $\boldsymbol{w}$, b y $\{\xi_n\}_{n=1}^N$ de $\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi})$, obtenemos la *representación dual* del problema de optimización (ecuación (9.7)), expresada por:

$$\begin{aligned} \widetilde{\boldsymbol{\alpha}} = & \arg\max_{\boldsymbol{\alpha}} \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m \boldsymbol{x}_n^\top \boldsymbol{x}_m \\ & \text{s.t. } 0 \leqslant \alpha_n \leqslant C \\ & \sum_{n=1}^{N} \alpha_n t_n = 0, \end{aligned} \tag{9.18}$$

$\forall n \in \{1, 2, \ldots, N\}$. La primera restricción esta formulada a partir de su naturaleza como multiplicadores de Lagrange, $\{\alpha_n \geqslant 0\}_{n=1}^N$, y para asegurar que $\{\mu_n \geqslant 0\}_{n=1}^N$ es evidente que $\{\alpha_n \leqslant C\}_{n=1}^N$ (ver ecuación (9.17)). Y la segunda condición es acorde a la derivada de $\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi})$ con respecto al parámetro b (ecuación (9.16)). Es importante recalcar que el problema de optimización dual es un *problema de programación cuadrático*–QPP de una sola variable, $\boldsymbol{\alpha}^*$.

Después de lo anterior, para clasificar nuevas muestras $\boldsymbol{x}_*$ usando el modelo entrenado evaluamos la función signo sobre $f(\boldsymbol{x}_*)$ definido en la ecuación (9.1), $\text{sign}(f(\boldsymbol{x}_*))$. El expresión del hiperplano, $f(\boldsymbol{x})$, puede ser expresado en termino de los $\{\widetilde{\alpha_n}\}_{n=1}^N$ (donde $\widetilde{\alpha_n}$ es el n-ésimo elemento de $\widetilde{\boldsymbol{\alpha}}$) y del producto interior entre la nueva muestra y las muestras de entrenamiento, sustituyendo $\boldsymbol{w}$ usando la ecuación (9.15), para obtener:

$$f(\boldsymbol{x}_*) = \sum_{n=1}^{N} \widetilde{\alpha}_n t_n \langle \boldsymbol{x}_*, \boldsymbol{x}_n \rangle + b \tag{9.19}$$

En el capítulo 8, mostramos que un problema de optimización restringida de esta forma satisface las condiciones *Karush-Huhn-Tucker*–(KKT), y se requiere que las propiedades

$$\alpha_n \geqslant 0 \tag{9.20}$$
$$t_n f(\boldsymbol{x}_n) - 1 + \xi_n \geqslant 0 \tag{9.21}$$
$$\alpha_n(t_n f(\boldsymbol{x}_n) - 1 + \xi_n) = 0, \tag{9.22}$$

se cumplan para todo n. Asimismo, para todas las muestras puede suceder que $\alpha_n{=}0$ ó $t_n f(\boldsymbol{x}_n){=}1$. Dicho lo anterior, cualquier muestra para la cual $\alpha_n = 0$, no contribuye a la sumatoria en la ecuación (9.19) y por ende tiene un papel en la predicción de nuevas muestras. Por otra parte, las muestras restantes, $\alpha_n{\neq}0$, son llamadas *vectores de soporte*–(SV), porque corresponden a las muestras que se encuentran sobre la margen generada y que se encuentran del lado equivocado del hiperplano (esto se ilustra en la figura 9.3); todas estas muestras satisfacen la condición $t_n f(\boldsymbol{x}_n){=}1 - \xi_n$. La propiedad anterior es central para la aplicabilidad de la SVM. Una vez el modelo es entrenado, una proporción significativa de las muestras pueden ser descartada y sólo los vectores de soporte son conservados.

Habiendo resulto el QPP y encontrado los valores de $\widetilde{\boldsymbol{\alpha}}$, podemos entonces determinar el valor del parámetro de sesgo b notando que cualquier vector de soporte $\boldsymbol{x}_n$ satisface $t_n f(\boldsymbol{x}_n){=}1$. Usando la ecuación (9.19) obtenemos:

$$t_n\left(\sum_{m\in\mathcal{S}} \widetilde{\alpha}_m t_m \langle \boldsymbol{x}_n, \boldsymbol{x}_m\rangle + b\right) = 1, \tag{9.23}$$

donde $\mathcal{S}$ es el conjunto de indices de los vectores de soporte. Aunque podemos solucionar esta ecuación para b usando un vector de soporte $\boldsymbol{x}_n$ escogido arbitrariamente, una solución más estable numéricamente es obtenida multiplicando primero por t_n, y como $t_n^2{=}1$, entonces despejamos b de la expresión; por último promediamos los resultados sobre todas las muestras que son vectores de soporte, obteniendo:

$$b = \frac{1}{N_{\mathcal{S}}}\sum_{n\in\mathcal{S}}\left(t_n - \sum_{m\in\mathcal{S}} \widetilde{\alpha}_m t_m \langle \boldsymbol{x}_n, \boldsymbol{x}_m\rangle\right), \tag{9.24}$$

donde $N_{\mathcal{S}}$ es el número de vectores de soporte total.

9.2.1. Máquinas de Vectores de Soporte en RKHS

Para extender las SVM a problema de clasificación no-lineales, se busca representar los datos en un espacio de alta dimensionalidad por medio de una función de mapeo no-lineal $\varphi(\cdot) : \mathbb{R}^P \to \mathbb{R}^D$, con $D \to \infty$. La idea es la misma a la del clasificador explicado anteriormente, generar un hiperplano de separación pero en un espacio de características de alta dimensionalidad ($\mathbb{R}^D$) en este orden de ideas, el hiperplano toma la forma:

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top\varphi(\boldsymbol{x}) + b{=}0. \tag{9.25}$$

Siguiendo el mismo procedimiento del apartado 9.2 y usando el mismo problema de optimización primario :

$$\begin{aligned} \widetilde{\boldsymbol{w}}, \widetilde{b}{=}&\underset{\boldsymbol{w},b,\boldsymbol{\xi}}{\arg\min}\frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{n=1}^{N}\xi_n \\ &\text{s.t. } t_n\Big(\boldsymbol{w}^\top\varphi(\boldsymbol{x}_n) + b\Big) \geqslant 1 - \xi_n, \\ &\xi_n \geqslant 0, \end{aligned} \tag{9.26}$$

para todo n, donde $\varphi(\boldsymbol{x}_n)\in\mathbb{R}^D$ es la función de mapeo no-lineal evaluada en la n-ésima muestra, $\boldsymbol{w}\in\mathbb{R}^D$ es el vector de pesos, ortogonal al hiperplano, y $b\in\mathbb{R}$ es el término de sesgo. Siguiendo el mismo procedimiento de la sección anterior y cambiando $\boldsymbol{x}_n$ por $\varphi(\boldsymbol{x}_n)$, y $\boldsymbol{x}_*$ por $\varphi(\boldsymbol{x}_*)$, llegamos al problema dual de la forma:

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}} =&\arg\max_{\boldsymbol{\alpha}} \sum_{n=1}^{N}\alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m t_n t_m \kappa(\boldsymbol{x}_n,\boldsymbol{x}_m)\,,\\ &\text{s.t. } 0\leqslant\alpha_n\leqslant C,\\ &\sum_{n=1}^{N}\alpha_n t_n=0,\end{aligned} \tag{9.27}$$

donde $\kappa:\mathbb{R}^P\times\mathbb{R}^P\to\mathbb{R}$ es una función kernel (ver apartado 7.2), con $\langle\varphi(\boldsymbol{x}_n)\,,\varphi(\boldsymbol{x}_m)\rangle=\kappa(\boldsymbol{x}_n,\boldsymbol{x}_m)$. De nuevo, el problema de optimización dual toma la forma de un QPP en función de solamente $\boldsymbol{\alpha}$.

En este orden de ideas, el hiperplano definido en la ecuación (9.25) en función de los $\{\widetilde{\alpha}_n\}_{n=N}^{N}$ para predecir una nueva muestra $\boldsymbol{x}_*$, toma la siguiente forma:

$$f(\boldsymbol{x}_*)=\sum_{n=1}^{N}\widetilde{\alpha}_n t_n\kappa(\boldsymbol{x}_*,\boldsymbol{x}_n)+b, \tag{9.28}$$

y la expresión para calcular el termino de sesgo b queda como:

$$b=\frac{1}{N_{\mathcal{S}}}\sum_{n\in\mathcal{S}}\left(t_n-\sum_{m\in\mathcal{S}}\widetilde{\alpha}_m t_m\kappa(\boldsymbol{x}_n,\boldsymbol{x}_m)\right). \tag{9.29}$$

Dicho todo lo anterior, es importante recordar que la solución de un QPP de P variables, en general tiene una complejidad computacional de $\mathcal{O}(P^3)$. Si vamos a la formulación dual a la que hemos llegado a partir del problema de optimización original, el cual involucra problemas de optimización (ver ecuaciones (9.7) y (9.26))) sobre variables de dimensión P, con N variables. Para un conjunto muestras fijas de dimensión P más pequeño que el número de muestras N, el procedimiento para llegar a una formulación dual parecería que posee desventajas. Sin embargo, permite al modelo ser reformulado usando kernels, por lo que el clasificador con margen máxima puede ser aplicado eficientemente en un espacio de características cuya dimensionalidad, $D\to\infty$, excede el número de muestras. La formulación del kernel también deja claro el rol de la restricción de la función kernel, tiene que ser definida positiva, ya que esta garantiza que el problema de optimización dual este bien definido.

9.2.2. Máquinas de Vectores de Soporte ponderadas (WSVM)

Debido que la formulación de la SVM no considera si la base de datos con la cual se desea entrenar es balanceada o no, puede ocurrir algún sesgo hacia alguna de las clases [6]. Además, dentro de varias aplicaciones del mundo real, una de las clases tiene una mayor penalización si el clasificador llega a equivocarse en ella en comparación a la otra clase. Por estos motivos, Osuna et al. en [55] proponen el clasificador *Weighted Support Vectors Machine*–(WSVM), el cual regulariza las variables auxiliares, $\{\xi_n\}$, dependiendo de la etiqueta, t_n. Dicho esto, el problema primal toma la forma:

$$\begin{aligned}\widetilde{\boldsymbol{w}}, \widetilde{b}=&\arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_+\sum_{t_n=+1}\xi_n + C_-\sum_{t_{n'}=-1}\xi_{n'}\\ &\text{s.t. } t_n\Big(\boldsymbol{w}^\top\boldsymbol{x}_n + b\Big) \geqslant 1-\xi_n,\\ &\xi_n \geqslant 0,\end{aligned} \tag{9.30}$$

donde $C_+, C_- \in \mathbb{R}^+$ los parámetros de regularización de las variables auxiliares para las muestras de la clase $+1$ y -1, respectivamente. Siguiendo el mismo procedimiento del apartado 9.2 llegamos al problema dual:

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}} =&\arg\max_{\boldsymbol{\alpha}} \sum_{n=1}^{N}\alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m t_n t_m \boldsymbol{x}_n^\top\boldsymbol{x}_m\\ &\text{s.t. } 0 \leqslant \alpha_n \leqslant C_+ \;\forall n \mid t_n=+1\\ &\quad 0 \leqslant \alpha_{n'} \leqslant C_- \;\forall n' \mid t_{n'}=-1\\ &\quad \sum_{n=1}^{N}\alpha_n t_n = 0;\end{aligned} \tag{9.31}$$

si suponemos un hiperplano en el espacio original de la forma de la ecuación (9.25), o si se supone un hiperplano en un espacio de características de alta dimensionalidad dado por la ecuación (9.25), el problema dual toma la forma:

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}} =&\arg\max_{\boldsymbol{\alpha}} \sum_{n=1}^{N}\alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m t_n t_m \kappa(\boldsymbol{x}_n,\boldsymbol{x}_m)\\ &\text{s.t. } 0 \leqslant \alpha_n \leqslant C_+ \;\forall n \mid t_n=+1\\ &\quad 0 \leqslant \alpha_{n'} \leqslant C_- \;\forall n' \mid t_{n'}=-1\\ &\quad \sum_{n=1}^{N}\alpha_n t_n = 0.\end{aligned} \tag{9.32}$$

La ecuación de decisión esta dada por la ecuación (9.19) o por la ecuación (9.28), dependiendo si se usa la SVM en el espacio original o en un espacio de características de alta dimensionalidad.

9.3. Máquinas de Vectores de Soporte Gemelas (TWSVM)

Desde su creación, las SVM han demostrado tener una gran capacidad de generalización y uso en diferentes aplicaciones de varias áreas del conocimiento. Aunque la investigación sobre las SVM ha logrado logros notables, todavía posee varias deficiencias en un estudio más profundo. Por ejemplo, los problemas que incluyen la relación entre la teoría de aprendizaje estadístico y otros sistemas teóricos, el procesamiento de grandes cantidades de datos y la selección de los parámetros. Especialmente, con el rápido desarrollo de la internet y de los sistemas de información, la alta dimensionalidad, la distribución y los datos dinámicos complejos son rápidamente generados.

Con el fin de reducir la complejidad computacional de la SVM, en 2006 Mangasarian y Wild Edward [56] propusieron un clasificador no-lineal, basados en la SVM, nombrada *generalized eigenvalue proximal support vector machine*–(GEPSVM). La idea principal del GEPSVM es buscar dos planos no-paralelos,

exigiendo que las muestras de cada clase estén lo más cerca a su correspondiente plano y a su vez lo más lejos posible de la otra clase. Aunque, la GEPSVM tiene un bajo tiempo de aprendizaje, su *acierto de clasificación* es bajo. En 2007, Jayadeva y Suresh [57] propusieron un nuevo método de aprendizaje de máquina llamado *twin support vector machine*–(TWSVM)—(máquinas de vectores de soporte gemelas, en español) para clasificación binaria con el espíritu del GEPSVM. El TWSVM busca generar dos planos no-paralelos, tal que cada hiperplano encierra a una de las dos clases, lo más lejos posible de la otra.

9.3.1. TWSVM lineal

Ahora, definimos las matrices $\boldsymbol{X}_+ \in \mathbb{R}^{P\times N_+} = [\boldsymbol{x}_i \mid t_i = +1]_{i=1}^{N_+}$ y $\boldsymbol{X}_- \in \mathbb{R}^{P\times N_-} = [\boldsymbol{x}_j \mid t_j = -1]_{i=1}^{N_-}$, donde $N = N_+ + N_-$, como las matrices de las clases $+1$ y -1, respectivamente. Observemos que hacemos uso de los indices $+$ y $-$ para hacer referencia a las clases $+1$ y -1, respectivamente. Como se dijo anteriormente, el TWSVM busca generar dos hiperplanos no-paralelos de la forma:

$$f_+(\boldsymbol{x}) = \boldsymbol{w}_+\boldsymbol{x} + b_+ = 0 \quad \text{y} \quad f_-(\boldsymbol{x}) = \boldsymbol{w}_-\boldsymbol{x} + b_- = 0, \tag{9.33}$$

donde $\boldsymbol{w}_+, \boldsymbol{w}_- \in \mathbb{R}^P$ son los vectores normales a sus respectivos hiperplanos y b_+,b_- son los términos de sesgo. Para encontrar los parámetros de ambos hiperplanos, dados en la ecuación (9.33), con el fin de minimizar la distancia de estos a las muestras correspondientes a su clase y que a su vez se encuentren lo más lejos posible de la otra clase, se busca resolver los siguientes problemas de optimización:

$$\begin{aligned} \widetilde{\boldsymbol{w}}_+, \widetilde{b}_+, \widetilde{\boldsymbol{\xi}}_- = & \underset{\boldsymbol{w}_+, b_+, \boldsymbol{\xi}_-}{\arg\min} \frac{1}{2}\|\boldsymbol{X}_+^\top \boldsymbol{w}_+ + \mathbf{1}_+ b_+\|_2^2 + c_{-,1}\mathbf{1}_-^\top \boldsymbol{\xi}_- \\ & \text{s.t.} \quad -1\Big(\boldsymbol{X}_-^\top \boldsymbol{w}_+ + \mathbf{1}_- b_+\Big) + \boldsymbol{\xi}_- \geqslant \mathbf{1}_- \\ & \boldsymbol{\xi}_- \geqslant 0, \end{aligned} \tag{9.34}$$

$$\begin{aligned} \widetilde{\boldsymbol{w}}_-, \widetilde{b}_-, \widetilde{\boldsymbol{\xi}}_+ = & \underset{\boldsymbol{w}_-, b_-, \boldsymbol{\xi}_+}{\arg\min} \frac{1}{2}\|\boldsymbol{X}_-^\top \boldsymbol{w}_- + \mathbf{1}_- b_-\|_2^2 + c_{+,1}\mathbf{1}_+^\top \boldsymbol{\xi}_+ \\ & \text{s.t.} \quad +1\Big(\boldsymbol{X}_+^\top \boldsymbol{w}_- + \mathbf{1}_+ b_-\Big) + \boldsymbol{\xi}_+ \geqslant \mathbf{1}_+ \\ & \boldsymbol{\xi}_- \geqslant 0, \end{aligned} \tag{9.35}$$

donde $\boldsymbol{\xi}_+ \in \mathbb{R}^{N_+}$, $\boldsymbol{\xi}_- \in \mathbb{R}^{N_-}$ son los vectores de variables auxiliares de las muestras de la clase $+1$ y -1 respectivamente, $c_{+,1}, c_{-,1} \in \mathbb{R}^+$ son los parámetros de regularización de las variables auxiliares; y, $\mathbf{1}_+ \in \mathbb{R}^{N_+}$ y $\mathbf{1}_- \in \mathbb{R}^{N_-}$ son vectores de unos.

El primer termino de las funciones objetivo, de las ecuaciones (9.34) y (9.35), es la suma al cuadrado del score de las muestras de la clase correspondiente al hiperplano, expresada por los hiperplanos definidos en la ecuación (9.33), la cual es proporcional a la suma al cuadrado de la distancia perpendicular del hiperplano a las muestras (recordar que la distancia perpendicular de un hiperplano a una muestra esta dado por $|t_n f_\ell(\boldsymbol{x})|/\|\boldsymbol{w}_\ell\|_2$). La restricciones requieren que el hiperplano tenga mínimo una distancia de 1 a las muestras de la otra clase; además, un conjunto de variables auxiliares son usadas para contrarrestar el inconveniente de que los hiperplanos no puedan tener una distancia de mínimo 1 con las muestras de la otra clase, es decir, con el fin de contrarrestar traslape entre las clases. El segundo termino es la suma de las variables auxiliares , con la meta de minimizar la mala clasificación. Esto se ilustra en la figura 9.4.

De ahora en adelante, con el fin de agilizar el proceso, usaremos las variables indice $\ell \in \{+, -\}$, $\ell' = -\ell$ y ζ_ℓ, donde $\zeta_\ell = +1$ si $\ell = +$ y -1 en caso contrario.

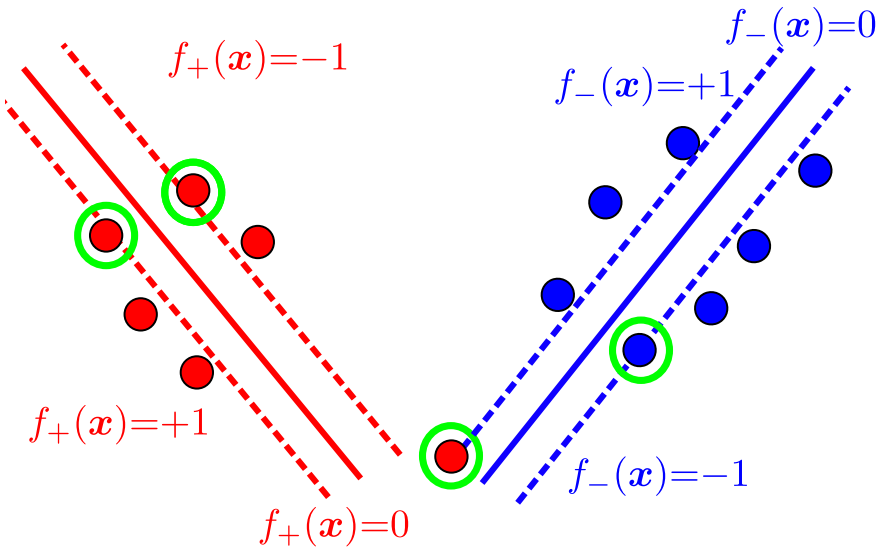


Figura 9.4.: La TWSVM genera dos hiperplanos no-paralelos, donde cada uno esta dedicado a una clase. Cada hiperplano debe estar lo más cerca a su clase correspondiente y lo más alejado posible de las muestras de la otra clase. Las muestras e hiperplanos de color rojo corresponden a la clase $+1$ y los de color azul a la clase -1. Las muestras que se encuentran cerradas en las circunferencias verdes, son las muestras que definen las margenes de los hiperplanos.

El Lagrangiano de los problemas primales expresados en las ecuaciones (9.34) y (9.35) están dados por:

$$\begin{aligned}\mathcal{L}_\ell(\boldsymbol{w}_\ell, b_\ell, \boldsymbol{\xi}_{\ell'}) = &\frac{1}{2}\Big(\boldsymbol{X}_\ell^\top \boldsymbol{w}_\ell + \mathbf{1}_\ell b_\ell\Big)^\top \Big(\boldsymbol{X}_\ell^\top \boldsymbol{w}_\ell + \mathbf{1}_\ell b_\ell\Big) + c_{\ell',1}\mathbf{1}_{\ell'} - \cdots \\ &\boldsymbol{\alpha}_\ell'^\top \Big(\zeta_{\ell'}\Big(\boldsymbol{X}_{\ell'}^\top \boldsymbol{w}_\ell + \mathbf{1}_{\ell'} b_\ell\Big) + \boldsymbol{\xi}_{\ell'} - \mathbf{1}_{\ell'}\Big) - \boldsymbol{\mu}_{\ell'}^\top \boldsymbol{\xi}_{\ell'}\end{aligned} \tag{9.36}$$

$\forall \ell \in \{+,-\}$, donde $\boldsymbol{\alpha}_{\ell'}, \boldsymbol{\mu}_{\ell'}, \mathbb{R}^{N_\ell'}$ son vectores con los valores de los multiplicadores de Lagrange. Su correspondiente conjunto de condiciones KKT son expresadas por:

$$\boldsymbol{\alpha}_{\ell'} \geqslant \mathbf{0}, \tag{9.37}$$

$$\zeta_{\ell'}\Big(\boldsymbol{X}_{\ell'}^\top \boldsymbol{w}_\ell + \mathbf{1}_{\ell'} b_\ell\Big) + \boldsymbol{\xi}_{\ell'} \geqslant \mathbf{1}_{\ell'}, \tag{9.38}$$

$$\boldsymbol{\alpha}_{\ell'}^\top \Big(\zeta_{\ell'}\Big(\boldsymbol{X}_{\ell'}^\top \boldsymbol{w}_\ell + \mathbf{1}_{\ell'} b_\ell\Big) + \boldsymbol{\xi}_{\ell'} - \mathbf{1}_{\ell'}\Big) = 0, \tag{9.39}$$

$$\boldsymbol{\mu}_{\ell'} \geqslant \mathbf{0}, \tag{9.40}$$

$$\boldsymbol{\xi}_{\ell'} \geqslant \mathbf{0}, \tag{9.41}$$

$$\boldsymbol{\mu}_{\ell'} \boldsymbol{\xi}_{\ell'} = 0, \tag{9.42}$$

$\forall \ell \in \{+,-\}$. Ahora hallamos las derivadas sobre $\boldsymbol{w}_\ell$, b_ℓ y $\boldsymbol{\xi}_{\ell'}$, como:

$$\frac{\partial \mathcal{L}_\ell}{\partial \boldsymbol{w}_\ell}=\mathbf{0} \Rightarrow \boldsymbol{X}_\ell \boldsymbol{X}_\ell^\top \boldsymbol{w}_\ell + \boldsymbol{X}_\ell \mathbf{1}_\ell b_\ell + \zeta_{\ell'} \boldsymbol{X}_{\ell'} \boldsymbol{\alpha}_{\ell'}=\mathbf{0}, \tag{9.43}$$

$$\frac{\partial \mathcal{L}_\ell}{\partial b_\ell}=0 \Rightarrow \mathbf{1}_\ell^\top \boldsymbol{X}_\ell^\top \boldsymbol{w}_\ell + \mathbf{1}_\ell^\top \mathbf{1}_\ell b_\ell + \zeta_{\ell'} \mathbf{1}_{\ell'}^\top \boldsymbol{\alpha}_{\ell'}=0, \tag{9.44}$$

$$\frac{\partial \mathcal{L}_\ell}{\partial \boldsymbol{\xi}_{\ell'}}=0 \Rightarrow c_{\ell',1} \mathbf{1}_{\ell'} - \boldsymbol{\alpha}_{\ell'} - \boldsymbol{\mu}_{\ell'}=0, \tag{9.45}$$

Luego, combinando las ecuaciones (9.43) y (9.44) llegamos a:

$$\left(\begin{bmatrix} \boldsymbol{X}_\ell \\ \mathbf{1}_\ell^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_\ell^\top & \mathbf{1}_\ell \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{w}_\ell \\ b_\ell \end{bmatrix} + \zeta_{\ell'} \begin{bmatrix} \boldsymbol{X}_{\ell'} \\ \mathbf{1}_{\ell'}^\top \end{bmatrix} \boldsymbol{\alpha}_{\ell'}=\mathbf{0}, \tag{9.46}$$

definiendo la matriz extendida $\boldsymbol{S}_\ell=\begin{bmatrix} \boldsymbol{X}_\ell^\top & \mathbf{1}_\ell \end{bmatrix}^\top$ y $\boldsymbol{z}_\ell=\begin{bmatrix} \boldsymbol{w}_\ell^\top & b_\ell \end{bmatrix}^\top$, y llevando la ecuación (9.46) en función de estas obtenemos que

$$\boldsymbol{S}_\ell^\top \boldsymbol{S}_\ell \boldsymbol{z}_\ell + \zeta_{\ell'} \boldsymbol{S}_{\ell'} \boldsymbol{\alpha}_{\ell'}=\mathbf{0} \quad \Rightarrow \quad \boldsymbol{z}_{\ell'}=\zeta_{\ell'} \Big(\boldsymbol{S}_\ell^\top \boldsymbol{S}_\ell \Big)^{-1} \boldsymbol{S}_{\ell'} \boldsymbol{\alpha}_{\ell'}. \tag{9.47}$$

$\forall \ell \in \{+,-\}$. Es importante notar que $\boldsymbol{S}_\ell^\top \boldsymbol{S}_\ell$ es siempre semi-definida positiva;sin embargo, es posible que no pueda estar bien condicionada en algunas ocasiones. Por lo anterior, se introduce un término de regularización del enfoque de regresión rígida tal como en [58], $\epsilon \boldsymbol{I}$ con $\epsilon \geqslant 0$, en aras de contrarrestar posibles problemas de mal condicionamiento de $\boldsymbol{S}_\ell^\top \boldsymbol{S}_\ell$. Aquí, $\boldsymbol{I}$ es una matriz identidad de tamaño necesario. Entonces introduciendo este termino de regularización, la ecuación (9.47) toma la forma:

$$\boldsymbol{z}_{\ell'}=\zeta_{\ell'} \Big(\boldsymbol{S}_\ell^\top \boldsymbol{S}_\ell + \epsilon \boldsymbol{I} \Big)^{-1} \boldsymbol{S}_{\ell'} \boldsymbol{\alpha}_{\ell'}, \tag{9.48}$$

Usando las definiciones de $\boldsymbol{S}_\ell$ y de $\boldsymbol{z}_\ell$ reescribimos el Lagrangiano en la ecuación ecuación (9.36) en función de estas dos variables. Luego, reemplazamos en esta la igualdad de la ecuación (9.47), si se requiere usar el termino de regularización ϵ por mal condicionamiento se reemplaza la ecuación (9.48), y obtenemos el problema de optimización dual:

$$\begin{aligned} \widetilde{\boldsymbol{\alpha}}_{\ell'} = \underset{\boldsymbol{\alpha}_{\ell'}}{\arg\max}\ & \mathbf{1}_{\ell'}^\top \boldsymbol{\alpha}_{\ell'} - \frac{1}{2} \boldsymbol{\alpha}_{\ell'}^\top \boldsymbol{S}_{\ell'}^\top \Big(\boldsymbol{S}_\ell \boldsymbol{S}_\ell^\top \Big)^{-1} \boldsymbol{S}_{\ell'}^\top \boldsymbol{\alpha}_{\ell'} \\ & \text{s.t. } \mathbf{0} \leqslant \boldsymbol{\alpha}_{\ell'} \leqslant c_{\ell',1}. \end{aligned} \tag{9.49}$$

Como $\boldsymbol{\mu}_{\ell'}, \boldsymbol{\xi}_{\ell'} \geqslant 0$ y teniendo en consideración la ecuación (9.45), tenemos la restricción de los problemas duales expresados en la ecuación (9.49). Podemos observar que los problemas duales toman la forma de dos QPPs, más pequeñas que la QPP del entrenamiento de la SVM. Observemos que cada una de las QPPs de la TWSVM tiene uan complejidadcomputacional de alrededor $\mathcal{O}(N_\ell^3)$ $\forall \ell \{+,-\}$. Supongamos que $N_+=N_-=N/2$, entonces tenemos que la complejidad computacional del TWSVM es de $\mathcal{O}\big(N^3/4\big)$ [57].

Ahora, una vez calculado los multiplicadores de Lagrange $\widetilde{\boldsymbol{\alpha}}_{\ell'}$ usamos la ecuación (9.47) ó 9.48, dependiendo del acondicionamiento de las matriz $\boldsymbol{S}_\ell \boldsymbol{S}_\ell^\top$. Con lo anterior, calculamos el vector $\boldsymbol{z}_\ell=\begin{bmatrix} \widetilde{\boldsymbol{w}}_\ell^\top & \widetilde{b}_\ell \end{bmatrix}^\top$, y extraemos los valores para $\widetilde{\boldsymbol{w}}_\ell$ y de $\widetilde{b}_\ell$ $\forall \ell \in \{+,-\}$. Es importante recalcar que para calcular, por ejemplo, los parámetros del hiperplano de la clase $+1$ es necesario encontrar los multiplicadores de Lagrange de la clase contraria, $\boldsymbol{\alpha}_-$. De lo anterior, podemos decir que los vectores de soporte de la clase ζ_ℓ provienen de las muestras de la clase $\zeta_{\ell'}$.

Por último, la etiqueta de una nueva muestra, $\boldsymbol{x}_*{\in}\mathbb{R}^P$, se asigna según a que hiperplano esta más cercana, haciendo uso de la distancia perpendicular. Lo anterior se resumen en la siguiente expresión como:

$$t_* = \underset{\ell\in\{+,-\}}{\arg\min} \frac{|f_\ell(\boldsymbol{x}_*)|}{\|\boldsymbol{w}_\ell\|_2}. \tag{9.50}$$

9.3.2. TWSVM No-lineal

La TWSVM extiende sus resultados a una extensión no-lineal, considerando la generación de dos superficies de separación en el espacio de entrada, construidas por alguna función kernel, en vez de hiperplanos. Las superficies de separación tienen la forma:

$$f_\ell(\boldsymbol{x}){=}\boldsymbol{k}(\boldsymbol{x})^\top\boldsymbol{w}_\ell + b_\ell{=}0, \quad \forall\ell{\in}\{+,-\}, \tag{9.51}$$

donde $\boldsymbol{k}(\cdot) : \mathbb{R}^P \to \mathbb{R}^N$ es un vector con elementos $k_n{=}\kappa(\boldsymbol{x}_n,\cdot)\ \forall n{\in}\{1,2,\ldots,N\}$, además $\boldsymbol{x}_n{\in}\widehat{\boldsymbol{X}}{=}[\boldsymbol{X}_+\ \boldsymbol{X}_-]$ y $\kappa(\cdot) : \mathbb{R}^P{\times}\mathbb{R}^P \to \mathbb{R}$ es una función kernel escogida con anticipación. Nótese, que para la versión no-lineal del TWSVM $\boldsymbol{w}_+, \boldsymbol{w}_-{\in}\mathbb{R}^N$. Dicho lo anterior, los problemas de optimización primales toman la forma:

$$\begin{aligned}
\widetilde{\boldsymbol{w}}_+, \widetilde{b}_+, \widetilde{\boldsymbol{\xi}}_- = &\underset{\boldsymbol{w}_+,b_+,\boldsymbol{\xi}_-}{\arg\min}\frac{1}{2}\|\boldsymbol{K}_+^\top\boldsymbol{w}_+ + \mathbf{1}_+b_+\|_2^2 + c_{-,1}\mathbf{1}_-^\top\boldsymbol{\xi}_- \\
&\text{s.t.} \quad -1\Big(\boldsymbol{K}_-^\top\boldsymbol{w}_+ + \mathbf{1}_-b_+\Big) + \boldsymbol{\xi}_- \geqslant \mathbf{1}_- \\
&\qquad \boldsymbol{\xi}_- \geqslant 0,
\end{aligned} \tag{9.52}$$

$$\begin{aligned}
\widetilde{\boldsymbol{w}}_-, \widetilde{b}_-, \widetilde{\boldsymbol{\xi}}_+ = &\underset{\boldsymbol{w}_-,b_-,\boldsymbol{\xi}_+}{\arg\min}\frac{1}{2}\|\boldsymbol{K}_-^\top\boldsymbol{w}_- + \mathbf{1}_-b_-\|_2^2 + c_{+,1}\mathbf{1}_+^\top\boldsymbol{\xi}_+ \\
&\text{s.t.} \quad +1\Big(\boldsymbol{K}_+^\top\boldsymbol{w}_- + \mathbf{1}_+b_-\Big) + \boldsymbol{\xi}_+ \geqslant \mathbf{1}_+ \\
&\qquad \boldsymbol{\xi}_- \geqslant 0,
\end{aligned} \tag{9.53}$$

donde $\boldsymbol{K}_\ell{\in}\mathbb{R}^{N\times N_\ell}$ con elementos $k_{nn'}{=}\kappa(\boldsymbol{x}_n,\boldsymbol{x}_m)$, tal que $\boldsymbol{x}_n{\in}\widehat{\boldsymbol{X}}$ y $\boldsymbol{x}_{n'}{\in}\boldsymbol{X}_\ell$. Siguiendo el mismo procedimiento que en el apartado 9.3.1 pero cambiando las matrices $\boldsymbol{X}_\ell\ \forall\ell{\in}\{+,-\}$ por las matrices $\boldsymbol{K}_\ell\ \forall\ell{\in}\{+,-\}$. En consecuencia, las matrices extendidas $\boldsymbol{S}_\ell{\in}\mathbb{R}^{(N+1)\times N_\ell}$ son calculadas de la forma $\boldsymbol{S}_\ell{=}\begin{bmatrix}\boldsymbol{K}_\ell^\top & \mathbf{1}_\ell\end{bmatrix}^\top$ $\forall\ell{\in}\{+,-\}$. Entonces, los problemas duales toman la misma forma que 9.49. Así mismo, la función de decisión de la ecuación (9.50) se reescribe como:

$$t_* = \underset{\ell\in\{+,-\}}{\arg\min}\frac{|f_\ell(\boldsymbol{x}_*)|}{\|\boldsymbol{K}^{1/2}\boldsymbol{w}_\ell\|_2}, \tag{9.54}$$

donde $\boldsymbol{K}{\in}\mathbb{R}^{N\times N}$ es la matriz de Gram de las muestras de entrenamiento, con elementos $k_{nn'}{=}\kappa(\boldsymbol{x}_n,\boldsymbol{x}_{n'})$; $n, n'{\in}\{1,2,\ldots N\}$ y $\boldsymbol{x}_n, \boldsymbol{x}_{n'}{\in}\widehat{\boldsymbol{X}}$.

9.4. Máquinas de Vectores de Soporte Gemelas Delimitadas (TBSVM)

Debido a que, la solución de los problemas duales del entrenamiento del TWSVM (ver ecuación (9.49)) dependen de que las matrices $\boldsymbol{S}_+\boldsymbol{S}_+^\top$ y $\boldsymbol{S}_-\boldsymbol{S}_-^\top$ sean invertibles, lo cual implicaría que se debe asumir que sean no-singulares. Sin embargo, este requisito no puede ser satisfecho siempre. Jayadeva en [57] propone

regularizar estas matrices por medio del termino $\epsilon \boldsymbol{I}$, $\epsilon \geqslant 0$, con el fin de contrarrestar el mal acondicionamiento de estas matrices. Aún así, la teoría dual en TWSVM no es perfecta [59]. Por lo anterior, Shao et al en [59] proponen algunas mejoras a la formulación del TWSVM, al cual nombraron *Twin Bounded Support Vector Machine*–(TBSVM). El TBSVM al igual que el TWSVM, busca construir dos hiperplanos no-paralelos, con la diferencia de que en los problemas de optimización (ver ecuaciones (9.34) y (9.35) y ecuaciones (9.52) y (9.53)) , el TBSVM agrega un término de regularización con la idea de maximizar la margen de los hiperplanos, por medio de minimizar $\|\boldsymbol{w}_\ell\|_2^2$ y b_ℓ^2 para todo $\ell \in \{+,-\}$.

9.4.1. TBSVM lineal

Al igual que la versión lineal del TWSVM, el TBSVM busca construir dos hiperplanos no-paralelos, ver ecuación (9.33). Los problemas primales que propone TBSVM están dados por:

$$\begin{aligned} \widetilde{\boldsymbol{w}}_\ell, \widetilde{b}_\ell, \widetilde{\boldsymbol{\xi}}_{\ell'} = & \underset{\boldsymbol{w}_+, b_+, \boldsymbol{\xi}_-}{\arg\min} \ \frac{c_{\ell,1}}{2}\Big(\|\boldsymbol{w}_\ell\|_2^2 + b_\ell^2\Big) + \frac{1}{2}\|\boldsymbol{X}_\ell^\top \boldsymbol{w}_\ell + \mathbf{1}_\ell b_\ell\|_2^2 + c_{\ell',2}\mathbf{1}_{\ell'}^\top \boldsymbol{\xi}_{\ell'} \\ & \text{s.t.} \quad \zeta_{\ell'}\Big(\boldsymbol{X}_{\ell'}^\top \boldsymbol{w}_\ell + \mathbf{1}_{\ell'} b_\ell\Big) + \boldsymbol{\xi}_{\ell'} \geqslant \mathbf{1}_{\ell'} \\ & \quad \boldsymbol{\xi}_{\ell'} \geqslant 0, \end{aligned} \tag{9.55}$$

$\forall \ell \in \{+,-\}$.Luego, el Lagrangiano para ambos problemas de optimización toma la forma:

$$\begin{aligned} \mathcal{L}_\ell(\boldsymbol{w}_\ell, b_\ell, \boldsymbol{\xi}_{\ell'}) = & \frac{c_{\ell,1}}{2}\Big(\boldsymbol{q}_\ell^\top \boldsymbol{w}_\ell + b_\ell^2\Big) + \frac{1}{2}\Big(\boldsymbol{X}_\ell^\top \boldsymbol{w}_\ell + \mathbf{1}_\ell b_\ell\Big)^\top \Big(\boldsymbol{X}_\ell^\top \boldsymbol{w}_\ell + \mathbf{1}_\ell b_\ell\Big) + c_{\ell',2}\mathbf{1}_{\ell'} - \cdots \\ & \boldsymbol{\alpha}_\ell'^\top \Big(\zeta_{\ell'}\Big(\boldsymbol{X}_{\ell'}^\top \boldsymbol{w}_\ell + \mathbf{1}_{\ell'} b_\ell\Big) + \boldsymbol{\xi}_{\ell'} - \mathbf{1}_{\ell'}\Big) - \boldsymbol{\mu}_{\ell'}^\top \boldsymbol{\xi}_{\ell'}, \end{aligned} \tag{9.56}$$

con $c_{\ell,1}, c_{\ell',2} \in \mathbb{R}^+$ como parámetros de regularización. El primer parámetro de regularización, penaliza el tamaño de los parámetros de los hiperplanos y el segundo regulariza las variables auxiliares, que indican cuando las muestras quedan mal clasificadas. El conjunto de condiciones KKT para los problemas de optimización del TBSVM son las mismos a las del TWSVM expresados en las ecuaciones (9.37) a (9.42). Ahora, hallamos las derivadas sobre $\boldsymbol{w}_\ell$, b_ℓ y $\boldsymbol{\xi}_{\ell'}$, de lo cual obtenemos:

$$\frac{\partial \mathcal{L}_\ell}{\partial \boldsymbol{w}_\ell} = \mathbf{0} \Rightarrow c_{\ell,1}\boldsymbol{w}_\ell + \boldsymbol{X}_\ell \boldsymbol{X}_\ell^\top \boldsymbol{w}_\ell + \boldsymbol{X}_\ell \mathbf{1}_\ell b_\ell + \zeta_{\ell'} \boldsymbol{X}_{\ell'} \boldsymbol{\alpha}_{\ell'} = \mathbf{0}, \tag{9.57}$$

$$\frac{\partial \mathcal{L}_\ell}{\partial b_\ell} = 0 \Rightarrow c_{\ell,1} b_\ell + \mathbf{1}_\ell^\top \boldsymbol{X}_\ell^\top \boldsymbol{w}_\ell + \mathbf{1}_\ell^\top \mathbf{1}_\ell b_\ell + \zeta_{\ell'} \mathbf{1}_{\ell'}^\top \boldsymbol{\alpha}_{\ell'} = 0, \tag{9.58}$$

$$\frac{\partial \mathcal{L}_\ell}{\partial \boldsymbol{\xi}_{\ell'}} = 0 \Rightarrow c_{\ell',2}\mathbf{1}_{\ell'} - \boldsymbol{\alpha}_{\ell'} - \boldsymbol{\mu}_{\ell'} = 0. \tag{9.59}$$

Siguiendo con el procedimiento, combinamos las ecuaciones (9.57) y (9.58) para obtener el siguiente sistema de ecuaciones:

$$\left(c_{\ell,1}\boldsymbol{I} + \begin{bmatrix} \boldsymbol{X}_\ell \\ \mathbf{1}_\ell^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_\ell^\top & \mathbf{1}_\ell \end{bmatrix}\right) \begin{bmatrix} \boldsymbol{w}_\ell \\ b_\ell \end{bmatrix} + \zeta_{\ell'} \begin{bmatrix} \boldsymbol{X}_{\ell'} \\ \mathbf{1}_{\ell'}^\top \end{bmatrix} \boldsymbol{\alpha}_{\ell'} = \mathbf{0}, \tag{9.60}$$

definimos la matriz extendida $\boldsymbol{S}_\ell = \begin{bmatrix} \boldsymbol{X}_\ell^\top & \mathbf{1}_\ell \end{bmatrix}^\top$ y $\boldsymbol{z}_\ell = \begin{bmatrix} \boldsymbol{w}_\ell^\top & b_\ell \end{bmatrix}^\top$, y reformulando la ecuación (9.60) en función de estas, y obtenemos:

$$\Big(c_{\ell,1}\boldsymbol{I} + \boldsymbol{S}_\ell^\top \boldsymbol{S}_\ell\Big)\boldsymbol{z}_\ell + \zeta_{\ell'}\boldsymbol{S}_{\ell'}\boldsymbol{\alpha}_{\ell'} = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{z}_{\ell'} = \zeta_{\ell'}\Big(c_{1,\ell}\boldsymbol{I} + \boldsymbol{S}_\ell^\top \boldsymbol{S}_\ell\Big)^{-1}\boldsymbol{S}_{\ell'}\boldsymbol{\alpha}_{\ell'}. \tag{9.61}$$

Usando las definiciones de $\boldsymbol{S}_\ell$ y de $\boldsymbol{z}_\ell$ reescribimos el Lagrangiano (de la ecuación (9.56)) en función de estas dos variables. Luego, reemplazamos en el Lagrangiano resultante la igualdad de la ecuación (9.61), y así se obtiene el problema de optimización dual de la forma:

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}}_{\ell'} = \underset{\boldsymbol{\alpha}_{\ell'}}{\arg\max}\ \mathbf{1}_{\ell'}^\top \boldsymbol{\alpha}_{\ell'} - \frac{1}{2}\boldsymbol{\alpha}_{\ell'}^\top \boldsymbol{S}_{\ell'}^\top \Big(c_{\ell,1}\boldsymbol{I} + \boldsymbol{S}_\ell \boldsymbol{S}_\ell^\top\Big)^{-1} \boldsymbol{S}_{\ell'}^\top \boldsymbol{\alpha}_{\ell'} \\ \text{s.t. } \mathbf{0} \leqslant \boldsymbol{\alpha}_{\ell'} \leqslant c_{\ell',2}.\end{aligned} \tag{9.62}$$

Igual que el TWSVM, para clasificar una nueva muestra $\boldsymbol{x}_*$ se tienen como criterio de decisión la mínima distancia entre el punto y los dos hiperplanos. Se usa la misma expresión de decisión que el TWSVM expresada en la ecuación (9.50).

9.4.2. TBSVM No-Lineal

De la misma forma que lo hizo las TWSVM, las TBSVM propuso una extensión no-lineal en donde en vez de construir dos hiperplanos, se generan dos superficies de separación generadas por una función kernel, y están expresadas por la ecuación 9.51.

Las funciones objetivos del TBSVM no-lineal son similares a las del TWSVM no-lineal, con la diferencia del termino de regularización. Dicho lo anterior, los problemas de optimización primales para el TBSVM no-lineal estan dados por:

$$\begin{aligned}\widetilde{\boldsymbol{w}}_\ell, \widetilde{b}_\ell, \widetilde{\boldsymbol{\xi}}_{\ell'} = &\underset{\boldsymbol{w}_\ell, b_\ell, \boldsymbol{\xi}_{\ell'}}{\arg\min} \frac{c_{\ell,1}}{2}\Big(\|\boldsymbol{w}_\ell\|_2^2 + b_\ell^2\Big) + \frac{1}{2}\|\boldsymbol{K}_\ell^\top \boldsymbol{w}_\ell + \mathbf{1}_\ell b_\ell\|_2^2 + c_{\ell',2}\mathbf{1}_{\ell'}^\top \boldsymbol{\xi}_{\ell'} \\ \text{s.t.} \quad & \zeta_{\ell'}\Big(\boldsymbol{K}_{\ell'}^\top \boldsymbol{w}_\ell + \mathbf{1}_{\ell'} b_\ell\Big) + \boldsymbol{\xi}_{\ell'} \geqslant \mathbf{1}_{\ell'} \\ & \boldsymbol{\xi}_{\ell'} \geqslant 0,\end{aligned} \tag{9.63}$$

$\forall \ell \in \{+,-\}$. Siguiendo el mismo procedimiento que en el apartado 9.4.1 con las mismas definiciones usadas en el apartado 9.3.2. Además, reemplazamos las matrices $\boldsymbol{X}_\ell$ por las matrices $\boldsymbol{K}_\ell$, $\forall \ell \in \{+,-\}$, previamente definidos en las secciones anteriores. Obtenemos las mismas formas de los problemas duales del TBSVM lineal, ver ecuación (9.62), con la excepción que la matrix extendida $\boldsymbol{S}_\ell = \big[\boldsymbol{K}_\ell^\top \quad \mathbf{1}_\ell\big]^\top$. Por último, la clasificación de una nueva muestra $\boldsymbol{x}_*$ se da por medio de la función de decisión 9.54.

9.5. Máquinas de vectores de soporte con un Lagrangiano ponderado

Como se ha expresado, las TWSVM y TBSVM tienen una complejidad computacional menor que las SVM estándar. Pero estos clasificadores poseen los mismos problemas ante bases de datos desbalanceadas. Como medida para contrarrestar estos problemas Shao et al. en [30] proponen *Weighted Lagragian Twin Support Vector Machine*–(WLTSVM), el cual al igual que las TWSVM busca construir dos hiperplanos no-paralelos. Además, ellos utilizan diferentes conjuntos de entrenamiento para generar ambos hiperplanos. El WLTSVM tiene como propiedades: (1) un submuestreo basado en grafos para reducir eficientemente el número de muestras de la clase de mayor muestras (clase mayoritaria), y el cual es robusto a muestras outliers; (2) la introducción análitica de un termino de sesgo ponderado,con el objetivo de mejorar el rendimiento sobre la clase de menor muestras (clase minoritaria), la cual tiende a ser ignorada por los clasificadores que no consideran el desbalance [6]; (3) los QPPs de este con respecto a los del TWSVM y TBSVM son diferentes, siendo más rápidos con rendimientos similares sobre base de datos sintéticas y del estado-del-arte.

9.5.1. WLTSVM lineal

Es importante recalcar que el WLTSVM fue diseñado para clasificación con bases de datos desbalanceadas, y por convención para estos problemas se toma la etiqueta $+1$ para la clase minoritaria (la de menor número de muestras) y la etiqueta -1 para la clase mayoritaria (la de mayor número de muestras). Dicho esto, la versión lineal del WLTSVM busca dos hiperplanos de la forma 9.33. Antes de definir el problema primal para el hiperplano minoritario (para la clase $+1$), debemos de definir el submuestreo basado en *ν-nearest neigbor*–(ν-NN), el cual busca caracterizar la densidad intra-clase de la clase mayoritaria (clase -1). En el grafo se considera que una muestra pertenece a la densidad intra-clase, cuando esta, junto a otra, son vecinos mutuos, ver figura 9.5(a). Esto implica que los puntos en las regiones de alta densidad tienen una mayor probabilidad de ser escogidos, mientras que los puntos que estén dentro de las regiones de baja densidad, por ejemplo, outliers, tendrán una baja probabilidad de ser seleccionados para el entrenamiento. Primero, definimos la matriz adyacente $\boldsymbol{U}$, con elementos:

$$U_{ij} = \begin{cases} \rho, & \text{Si } \boldsymbol{x}_i \in N_\nu(j) \text{ y } \boldsymbol{x}_j \in N_\nu(i); \\ 0, & \text{En otro caso,} \end{cases} \tag{9.64}$$

donde $\boldsymbol{x}_i, \boldsymbol{x}_j \in \boldsymbol{X}_-$, $N_\nu(j)$ denota el conjunto de ν-vecinos más cercanos, dentro de la clase mayoritaria, de la muestra $\boldsymbol{x}_j$, y $\rho \in \mathbb{R}^+$ es un escalar escogido con anticipación, $i, j \in \{1, 2, \ldots, N_-\}$. Entonces definimos el vector de coeficientes de submuestreo, $\boldsymbol{u}$ con elementos

$$u_i = \begin{cases} 1, & \text{Si } \sum_j U_{ij} \geqslant \nu; \\ 0, & \text{En otro caso.} \end{cases} \tag{9.65}$$

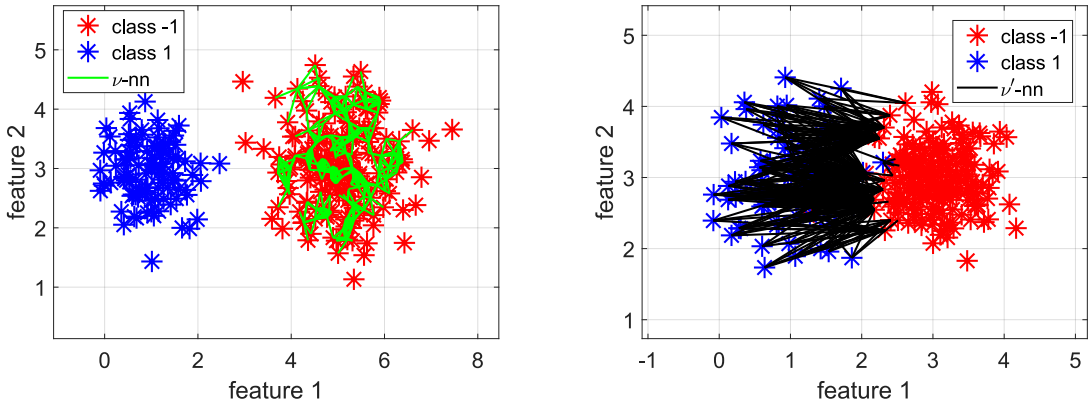
Las muestras $\boldsymbol{x}_i$ para la cual su correspondiente u_i es diferente de cero puede ser seleccionado, así obtenemos la matriz de muestras mayoritaria sub-muestreadas $\boldsymbol{X'}_- \in \mathbb{R}^{P \times N'_-}$, donde N'_- es el número de muestras de la clase mayoritaria escogidas por el submuestreo y $N_- \geqslant N'_-$. Fijémonos, que no hay forma de controlar el nivel de submuestreo, por esta razón es posible que no se elimine el desbalance entre ambas clases y hasta que llegue a suceder que $N_+ > N'_-$. De forma consecutiva, los autores introducen unos pesos en la construcción del hiperplano minoritario. Muy similar al WSVM, se define la matriz de pesos diagonal $\boldsymbol{D}_- = \text{mín}\{N_+/N'_-, N'_-/N_+\}\boldsymbol{I}$, con $\boldsymbol{D}_- \in \mathbb{R}^{N'_- \times N'_-}$.

Ahora, usando la matriz de muestras sub-muestreadas $\boldsymbol{X'}_-$ y la diagonal de pesos $\boldsymbol{D}_-$, encontraríamos los parámetros del hiperplano minoritario solucionando el problema de optimización:

$$\begin{aligned} \widetilde{\boldsymbol{w}}_+, \widetilde{b}_+, \widetilde{\boldsymbol{\xi}}_- = & \underset{\boldsymbol{w}_+, b_+, \boldsymbol{\xi}_-}{\arg\min} \frac{c_{+,1}}{2}\left(\|\boldsymbol{w}_+\|_2^2 + b_+^2 + \boldsymbol{\xi}_-^\top \boldsymbol{D}_- \boldsymbol{\xi}_-\right) + \frac{1}{2}\|\boldsymbol{X}_+^\top \boldsymbol{w}_+ + \boldsymbol{1}_+ b_+\|_2^2 \\ \text{s.t.} \quad & -1\left(\boldsymbol{X'}_-^\top \boldsymbol{w}_+ + \boldsymbol{1}_- b_+\right) + \boldsymbol{\xi}_- \geqslant \boldsymbol{1}_- \\ & \boldsymbol{\xi}_- \geqslant 0, \end{aligned} \tag{9.66}$$

donde $\boldsymbol{\xi}_- \in \mathbb{R}^{N'_-}$ son las variables auxiliares de las muestras de la clase mayoritaria sub-muestreadas y $c_{+,1} \in \mathbb{R}^+$.

Para el hiperplano mayoritario, se construye un grafo que caracterice la separabilidad inter-clase entre las muestras de la clase mayoritaria de la minoritaria. En el grafo inter-clase, un par de muestras son conectadas si ellas provienen de la clase minoritaria y mayoritaria, respectivamente, y la muestra de la clase mayoritaria es un ν'-NN de la muestra minoritaria (ver figura 9.5(b)). Esto implica que las muestras en la región marginal tienen más posibilidades de ser escogidos. Específicamente, la matriz adyacente al hiperplano mayoritario es denotada por $\boldsymbol{V} \in \mathbb{R}^{N_- \times N_+}$, con elementos:



(a) Intraclase (b) Interclase

Figura 9.5.: a) Densidad intraclase según el submuestreo basado en ν-nn. b) Densidad interclase según el submuestreo basado en ν'-nn. Para la generación de ambas gráficas fijamos $\nu{=}\nu'{=}6$.

$$V_{ij} = \begin{cases} \rho', & \text{Si } \boldsymbol{x}_i{\in}N_{\nu'}(j); \\ 0, & \text{En otro caso,} \end{cases} \tag{9.67}$$

donde $N_{\nu'}(j)$ denota el conjunto de muestras compuestas por los ν'-vecinos más cercanos, pertenecientes a la clase mayoritaria, de la muestra $\boldsymbol{x}_j{\in}\boldsymbol{X}_+$, y $\rho'{\in}\mathbb{R}^+$ es un escalar, $i{\in}\{1,2,\ldots,N_-\}$ y $j{\in}\{1,2,\ldots,N_+\}$. Entonces, se define el vector de coeficientes de submuestreo $\boldsymbol{v}{\in}\mathbb{R}^{N-}$ con elementos:

$$v_i = \begin{cases} 1, & \text{Si } \sum_j V_{ij} \geqslant \nu'; \\ 0, & \text{En otro caso.} \end{cases} \tag{9.68}$$

Luego construimos la matriz sub-muestreada para el hiperplano mayoritario $\bar{\boldsymbol{X}}_-{\in}\mathbb{R}^{P\times N'_-}$, $\bar{\boldsymbol{X}}_-{=}[\boldsymbol{x}_i \mid v_i{=}1]$, con $N'_- \leqslant N_-$ y $i{\in}1,2,\ldots,N_-$. De forma similar se define la matriz diagonal de pesos $\boldsymbol{D}_+{\in}\mathbb{R}^{N+\times N_+}$, dada por $\boldsymbol{D}_+{=}\text{mín}\{N_+/N'_-, N'_-/N_+\}\boldsymbol{I}$. Entonces, los parámetros del hiperplano mayoritarios se encontrarían resolviendo el problema de optimización:

$$\begin{aligned} \widetilde{\boldsymbol{w}}_-, \widetilde{b}_-, \widetilde{\boldsymbol{\xi}}_+ = & \underset{\boldsymbol{w}_-, b_-, \boldsymbol{\xi}_+}{\arg\min} \frac{c_{-,1}}{2}\left(\|\boldsymbol{w}_-\|_2^2 + b_-^2 + \boldsymbol{\xi}_+^\top \boldsymbol{D}_+ \boldsymbol{\xi}_+\right) + \frac{1}{2}\|\bar{\boldsymbol{X}}_-^\top \boldsymbol{w}_- + \boldsymbol{1}_- b_-\|_2^2 \\ \text{s.t.} \quad & \left(\boldsymbol{X}_+^\top \boldsymbol{w}_- + \boldsymbol{1}_+ b_-\right) + \boldsymbol{\xi}_+ \geqslant \boldsymbol{1}_+ \\ & \boldsymbol{\xi}_+ \geqslant \boldsymbol{0}, \end{aligned} \tag{9.69}$$

Para resolver los problemas de optimización expresados en 9.66 y 9.69, son llevados a una forma dual. Primero, el Lagrangiano de 9.66 esta dado por

$$\begin{aligned} \mathcal{L}_+(\boldsymbol{w}_+, b_+, \boldsymbol{\xi}_-, \boldsymbol{\alpha}_-){=}\frac{c_{+,1}}{2}\left(\boldsymbol{w}_+^\top \boldsymbol{w}_+ + b_+^2\right) + \frac{1}{2}\left(\boldsymbol{X}_+^\top \boldsymbol{w}_+ + \boldsymbol{1}_+ b_+\right)^\top \left(\boldsymbol{X}_+^\top \boldsymbol{w}_+ + \boldsymbol{1}_+ b_+\right) + \cdots \\ \cdots + c_{+,1}\boldsymbol{\xi}_-^\top \boldsymbol{D}_- \boldsymbol{\xi}_- - \boldsymbol{\alpha}_-^\top\left(-1\left(\boldsymbol{X}'^\top_- \boldsymbol{w}_+ + \boldsymbol{1}_- b_+\right) + \boldsymbol{\xi}_- - \boldsymbol{1}_-\right) \end{aligned} \tag{9.70}$$

donde $\boldsymbol{\alpha}_-{\in}\mathbb{R}^{N'}$ es el vector con los multiplicadores de Lagrange de las muestras sub-muestreadas para $\boldsymbol{X}'_-$. Las condiciones de KKT del problema 9.66 están dadas por:

$$-1\Big(\boldsymbol{X}_{-}^{\prime\top}\boldsymbol{w}_{+}+\mathbf{1}_{+}b_{+}\Big)+\boldsymbol{\xi}_{-}-\mathbf{1}_{-}\geqslant\mathbf{0} \tag{9.71}$$

$$\boldsymbol{\alpha}_{-}^{\top}\Big(-1\Big(\boldsymbol{X}_{-}^{\prime\top}\boldsymbol{w}_{+}+\mathbf{1}_{+}b_{+}\Big)+\boldsymbol{\xi}_{-}-\mathbf{1}_{-}\Big)=\mathbf{0} \tag{9.72}$$

$$\boldsymbol{\alpha}_{-}\geqslant\mathbf{0} \tag{9.73}$$

$$\boldsymbol{\xi}_{-}\geqslant\mathbf{0}. \tag{9.74}$$

Las derivadas con respecto a $\boldsymbol{w}_{+}$, b_{+} y $\boldsymbol{\xi}_{-}$ toman la forma:

$$\frac{\partial\mathcal{L}_{+}}{\partial\boldsymbol{w}_{+}}=\mathbf{0}\Rightarrow c_{+,1}\boldsymbol{w}_{+}+\boldsymbol{X}_{+}\boldsymbol{X}_{+}^{\top}\boldsymbol{w}_{+}+\boldsymbol{X}_{+}\mathbf{1}_{+}b_{+}+\boldsymbol{X}_{-}^{\prime}\boldsymbol{\alpha}_{-}=\mathbf{0} \tag{9.75}$$

$$\frac{\partial\mathcal{L}_{+}}{\partial b_{+}}=0\Rightarrow c_{+,1}b_{+}+\mathbf{1}_{+}^{\top}\boldsymbol{X}_{+}^{\top}\boldsymbol{w}_{+}+\mathbf{1}_{+}^{\top}\mathbf{1}_{+}b_{+}+\mathbf{1}_{-}^{\top}\boldsymbol{\alpha}_{-}=0 \tag{9.76}$$

$$\frac{\partial\mathcal{L}_{+}}{\partial\boldsymbol{\xi}_{-}}=\mathbf{0}\Rightarrow c_{+,1}\boldsymbol{D}_{-}\boldsymbol{\xi}_{-}-\boldsymbol{\alpha}_{-}=0. \tag{9.77}$$

Ahora, combinando las ecuaciones (9.75) y (9.76) nos conduce al siguiente sistema de ecuaciones, dado por:

$$\left(c_{+,1}\boldsymbol{I}+\begin{bmatrix}\boldsymbol{X}_{+}\\ \mathbf{1}_{+}^{\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_{+}^{\top} & \mathbf{1}_{+}\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{w}_{+}\\ b_{+}\end{bmatrix}+\begin{bmatrix}\boldsymbol{X}_{-}^{\prime}\\ \mathbf{1}_{-}^{\top}\end{bmatrix}\boldsymbol{\alpha}_{-}=\mathbf{0} \tag{9.78}$$

usando la definición de matriz extendida $\boldsymbol{S}_{+}$ usada en el apartado 9.3, definimos la matriz extendida $\boldsymbol{S}_{-}^{\prime}=\begin{bmatrix}\boldsymbol{X}_{-}^{\prime\top} & \mathbf{1}_{-}\end{bmatrix}^{\top}$ y $\boldsymbol{z}_{+}=\begin{bmatrix}\boldsymbol{w}_{+}^{\top} & b_{+}\end{bmatrix}^{\top}$. Usamos estas definiciones sobre la ecuación (9.78) y despejando $\boldsymbol{z}_{+}$ del resultado, hayamos que:

$$\boldsymbol{z}_{+}=-\Big(\boldsymbol{S}_{+}\boldsymbol{S}_{+}^{\top}+c_{+,1}\boldsymbol{I}\Big)^{-1}\boldsymbol{S}_{-}^{\prime}\boldsymbol{\alpha}_{-}. \tag{9.79}$$

Usando las definiciones antes mencionadas sobre el Lagrangeano 9.70 y sustituyendo la igualdad 9.79 en la misma, y usando la ecuación (9.77) obtenemos el problema dual de 9.66 como sigue:

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}}_{-}=&\arg\max_{\boldsymbol{\alpha}_{-}}-\frac{1}{2}\boldsymbol{\alpha}_{-}^{\top}\left(\boldsymbol{S}_{-}^{\prime\top}\Big(\boldsymbol{S}_{+}\boldsymbol{S}_{+}^{\top}+c_{+,1}\boldsymbol{I}\Big)^{-1}\boldsymbol{S}_{-}^{\prime}+\frac{1}{c_{+,1}}\boldsymbol{D}_{-}^{-1}\right)\boldsymbol{\alpha}_{-}+\mathbf{1}_{-}^{\top}\boldsymbol{\alpha}_{-}\\ &\text{s.t. }\boldsymbol{\alpha}_{-}\geqslant\mathbf{0}.\end{aligned} \tag{9.80}$$

Siguiendo el mismo procedimiento para obtener el problema dual del hiperplano de la clase minoritaria, resolvemos el problema de optimización 9.69 para entrenar el hiperplano de la clase mayoritaria, obteniendo:

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}}_{+}=&\arg\max_{\boldsymbol{\alpha}_{+}}-\frac{1}{2}\boldsymbol{\alpha}_{+}^{\top}\left(\boldsymbol{S}_{+}^{\top}\Big(\bar{\boldsymbol{S}}_{-}\bar{\boldsymbol{S}}_{-}^{\top}+c_{-,1}\boldsymbol{I}\Big)^{-1}\boldsymbol{S}_{+}+\frac{1}{c_{-,1}}\boldsymbol{D}_{+}^{-1}\right)\boldsymbol{\alpha}_{+}+\mathbf{1}_{+}^{\top}\boldsymbol{\alpha}_{+}\\ &\text{s.t. }\boldsymbol{\alpha}_{+}\geqslant\mathbf{0}.\end{aligned} \tag{9.81}$$

donde $\boldsymbol{\alpha}_{+}\in\mathbb{R}^{N_{+}}$ es el vector con los multiplicadores de Lagrange de las muestras de la clase $+1$ (minoritarias), $\bar{\boldsymbol{S}}_{-}=\begin{bmatrix}\bar{\boldsymbol{X}}_{-}^{\top} & \mathbf{1}_{-}\end{bmatrix}^{\top}$ y $\boldsymbol{z}_{-}\begin{bmatrix}\boldsymbol{w}_{-}^{\top} & b_{-}\end{bmatrix}^{\top}$; a partir de las condiciones de KKT obtenemos que:

$$\boldsymbol{z}_{-}=\Big(\bar{\boldsymbol{S}}_{-}\bar{\boldsymbol{S}}_{-}^{\top}+c_{-,1}\boldsymbol{I}\Big)^{-1}\boldsymbol{S}_{+}\boldsymbol{\alpha}_{+}. \tag{9.82}$$

Una vez se haya solucionado $\boldsymbol{w}_{+}$, b_{+}, $\boldsymbol{w}_{-}$ y b_{-}, una nueva muestra $\boldsymbol{x}_{*}\mathbb{R}^{P}$ es clasificada por medio de la función de decisión 9.50.

9.5.2. WLTSVM No-Lineal

Al igual que el TWSVM y el TBSVM, el WLTSVM realiza una extensión para clasificación no-lineal, a través de la creación de superficies de separación generadas por una función kernel escogida con antelación. Estas superficies $f_+(\boldsymbol{x})=0$ y $f_-(\boldsymbol{x})=0$ están definidas en la ecuación (9.51).

Siguiendo con el mismo procedimiento que en su versión lineal y habiendo generado las matrices submuestreadas $\boldsymbol{X'}_-$ y $\bar{\boldsymbol{X}}_-$. Y definiendo la matriz kernel $\boldsymbol{K'}_-\in\mathbb{R}^{N\times N'_-}$ con elementos $k_{nn'}=\kappa(\boldsymbol{x}_n,\boldsymbol{x}_{n'})$ donde $\boldsymbol{x}_n\in\widehat{\boldsymbol{X}}$ y $\boldsymbol{x}_{n'}\in\boldsymbol{X'}_-$; y $\bar{\boldsymbol{K}}_-\in\mathbb{R}^{N\times N''_-}$ con elementos $k_{nn'}=\kappa(\boldsymbol{x}_n,\boldsymbol{x}_{n'})$ donde $\boldsymbol{x}_n\in\widehat{\boldsymbol{X}}$ y $\boldsymbol{x}_{n'}\in\bar{\boldsymbol{X}}_-$.
Los problemas de optimización toman la forma:

$$\begin{aligned}\widetilde{\boldsymbol{w}}_+,\widetilde{b}_+,\widetilde{\boldsymbol{\xi}}_- =& \underset{\boldsymbol{w}_+,b_+,\boldsymbol{\xi}_-}{\arg\min}\ \frac{c_{+,1}}{2}\Big(\|\boldsymbol{w}_+\|_2^2+b_+^2+\boldsymbol{\xi}_-^\top\boldsymbol{D}_-\boldsymbol{\xi}_-\Big)+\frac{1}{2}\|\boldsymbol{K}_+^\top\boldsymbol{w}_++\boldsymbol{1}_+b_+\|_2^2\\ &\text{s.t.}\quad -1\Big(\boldsymbol{K'}_-^\top\boldsymbol{w}_++\boldsymbol{1}_-b_+\Big)+\boldsymbol{\xi}_-\geqslant\boldsymbol{1}_-\\ &\qquad\boldsymbol{\xi}_-\geqslant 0,\end{aligned}\tag{9.83}$$

$$\begin{aligned}\widetilde{\boldsymbol{w}}_-,\widetilde{b}_-,\widetilde{\boldsymbol{\xi}}_+ =& \underset{\boldsymbol{w}_-,b_-,\boldsymbol{\xi}_+}{\arg\min}\ \frac{c_{-,1}}{2}\Big(\|\boldsymbol{w}_-\|_2^2+b_-^2+\boldsymbol{\xi}_+^\top\boldsymbol{D}_+\boldsymbol{\xi}_+\Big)+\frac{1}{2}\|\bar{\boldsymbol{K}}_-^\top\boldsymbol{w}_-+\boldsymbol{1}_-b_-\|_2^2\\ &\text{s.t.}\quad \Big(\boldsymbol{K}_+^\top\boldsymbol{w}_-+\boldsymbol{1}_+b_-\Big)+\boldsymbol{\xi}_+\geqslant\boldsymbol{1}_+\\ &\qquad\boldsymbol{\xi}_+\geqslant\boldsymbol{0},\end{aligned}\tag{9.84}$$

Siguiendo el mismo procedimiento que en el apartado 9.5.1 podemos llegar a las mismas formas duales que en 9.80 y 9.81, con las matrices extendidas $\boldsymbol{S}_+=\begin{bmatrix}\boldsymbol{K}_+^\top & \boldsymbol{1}_+\end{bmatrix}^\top$, $\boldsymbol{S'}_-=\begin{bmatrix}\boldsymbol{K'}_-^\top & \boldsymbol{1}_-\end{bmatrix}^\top$ y $\bar{\boldsymbol{S}}_-=\begin{bmatrix}\bar{\boldsymbol{K}}_-^\top & \boldsymbol{1}_-\end{bmatrix}^\top$. También los vectores $\boldsymbol{z}_+$ y $\boldsymbol{z}_-$ toman la misma forma que las ecuaciones (9.79) y (9.82), respectivamente. La clasificación de una nueva muestra $\boldsymbol{x}_*$ esta dada por la función de decisión 9.54.

Parte II.

Propuesta

10. Máquinas de Vectores de Soporte Gemelas mejorada (ETWSVM)

Inducimos la solución de dos problemas claves que surgen cuando se tratan con datos desbalanceados dentro de modelos basados en TWSVM: *i)* la extensión no-lineal de la función de decisión a través de un mapeo adecuado a *espacios de Hilbert con kernel reproductivo*–(RKHS) con el fin de mejorar la representación de los datos de entrada; y *ii)* la estimación de los parámetros de la función kernel, por ejemplo, la matriz de covarianza de un kernel Gaussiano, para aprovechar la separabilidad de los datos y para evitar un conocimiento previo del usuario con respecto al ajuste de la función kernel.

Presentamos *máquinas de vectores de soporte gemelas mejorada*–(ETWSVM) para la clasificación binaria de datos desbalanceadas. Nuestro clasificador comprende una reforma a la formulación dual del problema de optimización cuadrática del TBSVM basándose del bien conocido lema de la matriz inversa, lo cual permite codificar una función de decisión en un RKHS con una dimensión posiblemente infinita. Además, una técnica basada en alineamiento de kernel es incluido durante el entrenamiento del ETWSVM como una estimación basada en datos de la función de similitud, kernel. Por último, nuestra propuesta es extendida a problemas multi-clase a través de las estrategias *uno-versus-Resto*–(OvR) y *uno-versus-uno*–(OvO) [39].

10.1. Fundamentos en ETWSVM

Nuestra propuesta básicamente es una reformulación de la versión no-lineal del TBSVM. Comenzando que las versiones no-lineales del TWSVM y del TBSVM buscan dos superficies de separación. Sugerimos que en vez de estas dos superficies se busque la construcción de dos hiperplanos no-paralelos en un espacio de RKHS de la formas

$$f_+(\boldsymbol{x}) = \varphi(\boldsymbol{x})^\top \boldsymbol{w}_+ + b_+ \quad \text{y} \quad f_-(\boldsymbol{x}) = \varphi(\boldsymbol{x})^\top \boldsymbol{w}_- + b_- \tag{10.1}$$

donde $\boldsymbol{w}_\ell \in \mathcal{H}$ es el el vector normal al hiperplano correspondiente, $b_\ell \in \mathbb{R}$ es el termino de sesgo del hiperplano correspondiente, para todo $\ell \in \{+,-\}$, y $\mathcal{H} \subset \mathbb{R}^Q$ es un RKHS. En general, $\mathcal{H}$ se puede pensar como un espacio de características de alta-dimensionalidad (de "infinita" dimensionalidad, $Q \to \infty$) [11]. Al igual que el TBSVM versión lineal, apartado 9.4, se busca que cada hiperplano este lo más cerca posible de las muestras de su respectiva clase y lo más lejos posible de la otra clase y a la vez regularizar los parámetros de los hiperplanos. Dicho lo anterior, los problemas de optimización primal están dadas por:

$$\begin{aligned} \widetilde{\boldsymbol{w}}_\ell, \widetilde{b}_\ell, \widetilde{\boldsymbol{\xi}}_{\ell'} = & \underset{\boldsymbol{w}_+, b_+, \boldsymbol{\xi}_-}{\arg\min} \ \frac{c_{\ell,1}}{2}\Big(\|\boldsymbol{w}_\ell\|_2^2 + b_\ell^2\Big) + \frac{1}{2}\|\boldsymbol{\Phi}_\ell^\top \boldsymbol{w}_\ell + \mathbf{1}_\ell b_\ell\|_2^2 + c_{\ell',2}\mathbf{1}_{\ell'}^\top \boldsymbol{\xi}_{\ell'} \\ & \text{s.t.} \quad \zeta_{\ell'}\Big(\boldsymbol{\Phi}_{\ell'}^\top \boldsymbol{w}_\ell + \mathbf{1}_{\ell'} b_\ell\Big) + \boldsymbol{\xi}_{\ell'} \geqslant \mathbf{1}_{\ell'} \\ & \quad \boldsymbol{\xi}_{\ell'} \geqslant 0, \end{aligned} \tag{10.2}$$

para $\ell\in\{+,-\}$; donde $\boldsymbol{\Phi}_\ell\in\mathbb{R}^{Q\times N_\ell}$ es la matriz de entrada con las muestras de la clase ζ_ℓ mapeadas por medio de la función de mapeo $\varphi(\cdot):\mathbb{R}^P\to\mathcal{H}$. Es pertinente recordar que $\zeta_\ell{=}{+}1$ si $\ell{=}{+}$, y $\zeta_\ell{=}{-}1$ para el caso contrario. Dicho lo anterior, podemos definir el Lagrangiano para calcular el hiperplano $f_\ell(\boldsymbol{x}){=}0$ como:

$$\begin{aligned}\mathcal{L}_\ell(\boldsymbol{w}_\ell,b_\ell,\boldsymbol{\xi}_{\ell'},\boldsymbol{\alpha}_{\ell'},\boldsymbol{\mu}_{\ell'}){=}\frac{c_{\ell,1}}{2}\Big(\boldsymbol{w}_\ell^\top\boldsymbol{w}_\ell+b_\ell^2\Big)+\frac{1}{2}\Big(\boldsymbol{\Phi}_\ell^\top\boldsymbol{w}_\ell+\mathbf{1}_\ell b_\ell\Big)^\top\Big(\boldsymbol{\Phi}_\ell^\top\boldsymbol{w}_\ell+\mathbf{1}_\ell b_\ell\Big)+\cdots\\ \cdots+c_{2,\ell'}\mathbf{1}_{\ell'}^\top\boldsymbol{\xi}_{\ell'}-\boldsymbol{\alpha}_{\ell'}^\top\Big(\zeta_{\ell'}\Big(\boldsymbol{\Phi}_{\ell'}^\top\boldsymbol{w}_\ell+\mathbf{1}_{\ell'}b_\ell\Big)+\boldsymbol{\xi}_{\ell'}-\mathbf{1}_{\ell'}\Big)-\boldsymbol{\mu}_{\ell'}^\top\boldsymbol{\xi}_{\ell'},\end{aligned}\tag{10.3}$$

donde $\boldsymbol{\xi}_{\ell'},\boldsymbol{\mu}_{\ell'}\in\mathbb{R}^{N_{\ell'}}$ son los vectores con los multiplicadores de Lagrange de las condiciones expresadas en el problema de optimización 10.2 para las muestras de la clase $\zeta_{\ell'}$. Las condiciones KKT del problema de optimización están dados por:

$$\boldsymbol{\alpha}_{\ell'}\geqslant\mathbf{0},\tag{10.4}$$

$$\zeta_{\ell'}\Big(\boldsymbol{\Phi}_{\ell'}^\top\boldsymbol{w}_\ell+\mathbf{1}_{\ell'}b_\ell\Big)+\boldsymbol{\xi}_{\ell'}\geqslant\mathbf{1}_{\ell'},\tag{10.5}$$

$$\boldsymbol{\alpha}_{\ell'}^\top\Big(\zeta_{\ell'}\Big(\boldsymbol{\Phi}_{\ell'}^\top\boldsymbol{w}_\ell+\mathbf{1}_{\ell'}b_\ell\Big)+\boldsymbol{\xi}_{\ell'}-\mathbf{1}_{\ell'}\Big)=0,\tag{10.6}$$

$$\boldsymbol{\mu}_{\ell'}\geqslant\mathbf{0},\tag{10.7}$$

$$\boldsymbol{\xi}_{\ell'}\geqslant\mathbf{0},\tag{10.8}$$

$$\boldsymbol{\mu}_{\ell'}\boldsymbol{\xi}_{\ell'}{=}0.\tag{10.9}$$

Luego, las derivadas con respecto a $\boldsymbol{w}_\ell$, b_ℓ y $\boldsymbol{\xi}_{\ell'}$ toman la forma:

$$\frac{\partial\mathcal{L}_\ell}{\partial\boldsymbol{w}_\ell}{=}\mathbf{0}\Rightarrow c_{\ell,1}\boldsymbol{w}_\ell+\boldsymbol{\Phi}_\ell\boldsymbol{\Phi}_\ell^\top\boldsymbol{w}_\ell+\boldsymbol{\Phi}_\ell\mathbf{1}_\ell b_\ell+\zeta_{\ell'}\boldsymbol{\Phi}_{\ell'}\boldsymbol{\alpha}_{\ell'}{=}\mathbf{0},\tag{10.10}$$

$$\frac{\partial\mathcal{L}_\ell}{\partial b_\ell}{=}0\Rightarrow c_{\ell,1}b_\ell+\mathbf{1}_\ell^\top\boldsymbol{\Phi}_\ell^\top\boldsymbol{w}_\ell+\mathbf{1}_\ell^\top\mathbf{1}_\ell b_\ell+\zeta_{\ell'}\mathbf{1}_{\ell'}^\top\boldsymbol{\alpha}_{\ell'}{=}0,\tag{10.11}$$

$$\frac{\partial\mathcal{L}_\ell}{\partial\boldsymbol{\xi}_{\ell'}}{=}0\Rightarrow c_{\ell',2}\mathbf{1}_{\ell'}-\boldsymbol{\alpha}_{\ell'}-\boldsymbol{\mu}_{\ell'}{=}0.\tag{10.12}$$

Combinando las ecuaciones (10.10) y (10.11) con el fin de generar el sistema de ecuaciones,

$$\left(c_{\ell,1}\boldsymbol{I}+\begin{bmatrix}\boldsymbol{\Phi}_\ell\\ \mathbf{1}_\ell^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\Phi}_\ell^\top & \mathbf{1}_\ell\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{w}_\ell\\ b_\ell\end{bmatrix}+\zeta_{\ell'}\begin{bmatrix}\boldsymbol{\Phi}_{\ell'}\\ \mathbf{1}_{\ell'}^\top\end{bmatrix}\boldsymbol{\alpha}_{\ell'}{=}\mathbf{0},\tag{10.13}$$

y definiendo las matrices extendidas $\boldsymbol{S}_\ell{=}\begin{bmatrix}\boldsymbol{\Phi}_\ell^\top & \mathbf{1}_\ell\end{bmatrix}^\top$ y $\boldsymbol{z}_\ell{=}\begin{bmatrix}\boldsymbol{w}_\ell^\top & b_\ell\end{bmatrix}^\top$, para $\ell\in\{+,-\}$. Haciendo uso de las matrices definidas anteriormente, el sistema de ecuaciones 10.13 se reescribe de la siguiente forma:

$$\Big(c_{\ell,1}\boldsymbol{I}+\boldsymbol{S}_\ell\boldsymbol{S}_\ell^\top\Big)\boldsymbol{z}_\ell-\zeta_{\ell'}\boldsymbol{S}_{\ell'}\boldsymbol{\alpha}_{\ell'}{=}\mathbf{0}.\tag{10.14}$$

Ahora, despejamos $\boldsymbol{z}_\ell$ de la ecuación (10.14), y obtenemos que:

$$\boldsymbol{z}_\ell{=}\zeta_{\ell'}\Big(c_{\ell,1}\boldsymbol{I}+\boldsymbol{S}_\ell\boldsymbol{S}_\ell^\top\Big)^{-1}\boldsymbol{S}_{\ell'}\boldsymbol{\alpha}_{\ell'}.\tag{10.15}$$

Notemos que el calculo de la matrices de covarianza $\boldsymbol{\Phi}_\ell\boldsymbol{\Phi}_\ell^\top$, para $\ell\in\{+,-\}$, están en el espacio RKHS, las cuales no pueden ser calculadas debido a la "infinita"-dimensionalidad, propiedad de la función de mapeo $\varphi(\cdot)$. Sin embargo, aplicamos el lema de la matriz inversa, dado por:

$$(\boldsymbol{A}+\boldsymbol{BCD})^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}\big(\boldsymbol{C}^{-1}+\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B}\big)^{-1}\boldsymbol{D}\boldsymbol{A}^{-1}, \tag{10.16}$$

y haciendo uso de las identificaciones:

$$c_{\ell,1}\boldsymbol{I}\rightarrow\boldsymbol{A}\quad \boldsymbol{S}_\ell\rightarrow\boldsymbol{B}\quad \boldsymbol{I}\rightarrow\boldsymbol{C}\quad \boldsymbol{S}_\ell^\top\rightarrow\boldsymbol{D} \tag{10.17}$$

obtenemos que:

$$\Big(c_{\ell,1}\boldsymbol{I}+\boldsymbol{S}_\ell\boldsymbol{S}_\ell^\top\Big)^{-1}=\frac{1}{2c_{\ell,1}}\Big(\boldsymbol{I}-\boldsymbol{S}_\ell(c_{\ell,1}\boldsymbol{I}+\boldsymbol{S}_\ell\boldsymbol{S}_\ell)^{-1}\boldsymbol{S}_\ell^\top\boldsymbol{S}_{\ell'}\Big)\boldsymbol{\alpha}_{\ell'}, \tag{10.18}$$

para $\ell\in\{+,-\}$. Dicho lo anterior, llevamos el Lagrangiano (ecuación (10.3)) en función de las matrices $\boldsymbol{S}_\ell$ y del vector $\boldsymbol{z}_\ell$, para $\ell\in\{+,-\}$, y reemplazando este último ($\boldsymbol{z}_\ell$) acorde a la ecuación (10.15), lo cual nos lleva a que el problema de optimización dual para el ETWSVM toma la forma:

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\ell'}=&\underset{\boldsymbol{\alpha}_{\ell'}}{\arg\max}\ \mathbf{1}_{\ell'}^\top\boldsymbol{\alpha}_{\ell'}-\frac{1}{2c_{\ell,1}}\boldsymbol{\alpha}_{\ell'}^\top\Big(\boldsymbol{S}_{\ell'}^\top\boldsymbol{S}_\ell'-\boldsymbol{S}_{\ell'}^\top\boldsymbol{S}_\ell\Big(c_{\ell,1}\boldsymbol{I}+\boldsymbol{S}_\ell^\top\boldsymbol{S}_\ell\Big)^{-1}\boldsymbol{S}_\ell^\top\boldsymbol{S}_{\ell'}\Big)\boldsymbol{\alpha}_{\ell'}^\top\\ &\text{s.t. }\mathbf{0}\leqslant\boldsymbol{\alpha}_{\ell'}\leqslant c_{\ell',2}\mathbf{1}_{\ell'}.\end{aligned} \tag{10.19}$$

Basándonos en el producto interno en RKHS, las matrices $\boldsymbol{K}_{\ell,\ell'}$, $\hat{\boldsymbol{K}}_{\ell,\ell'}\in\mathbb{R}^{N_\ell\times N_{\ell'}}$ pueden ser inferidas como: $\boldsymbol{K}_{\ell,\ell'}=\boldsymbol{\Phi}_\ell^\top\boldsymbol{\Phi}_{\ell'}$ y $\hat{\boldsymbol{K}}_{\ell,\ell'}=\boldsymbol{S}_\ell^\top\boldsymbol{S}_{\ell'}=\hat{\boldsymbol{K}}_{\ell,\ell'}+\mathbf{1}_\ell\mathbf{1}_{\ell'}^\top$, respectivamente. Es importante precisar que la matriz $\boldsymbol{K}_{\ell,\ell'}$ tiene elementos $k_{nn'}=\kappa(\boldsymbol{x}_n,\boldsymbol{x}_{n'})=\langle\varphi(\boldsymbol{x}_n)\,,\varphi(\boldsymbol{x}_{n'})\rangle_{\mathcal{H}}$, con $n\in\{1,\ldots,N_\ell\}$ y $n'\in\{1,\ldots,N_{\ell'}\}$, haciendo uso de lo expresado en el apartado 7.2, donde $\kappa:\mathbb{R}^P\times\mathbb{R}^P\rightarrow\mathbb{R}$ es la función kernel escogida con anticipación. En este sentido, el problema de optimización dual de la ecuación (10.19) queda de la forma

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\ell'}=&\underset{\boldsymbol{\alpha}_{\ell'}}{\arg\max}\ \mathbf{1}_{\ell'}^\top\boldsymbol{\alpha}_{\ell'}-\frac{1}{2c_{\ell',2}}\boldsymbol{\alpha}_{\ell'}^\top\Big(\hat{\boldsymbol{K}}_{\ell',\ell'}-\hat{\boldsymbol{K}}_{\ell',\ell}\boldsymbol{A}_\ell^{-1}\hat{\boldsymbol{K}}_{\ell,\ell'}\Big)\boldsymbol{\alpha}_{\ell'}\\ &\text{s.t. }\mathbf{0}\leqslant\boldsymbol{\alpha}_{\ell'}\leqslant c_{\ell',2}\mathbf{1}_{\ell'},\end{aligned} \tag{10.20}$$

donde $\boldsymbol{A}_\ell=\hat{\boldsymbol{K}}_{\ell,\ell}+c_{\ell,1}\boldsymbol{I}$. La solución de los problemas de programación cuadrática en 10.20 (uno por cada $\boldsymbol{\alpha}_{\ell'}$, $\forall\ell'\in\{+,-\}$), permiten obtener los vectores de pesos $\boldsymbol{z}_\ell\in\mathbb{R}^{Q+1}$ en $\mathcal{H}$, El cual puede ser expresado como en la ecuación (10.15). De nuevo, el calculo de la matriz de covarianza en RKHS es eludido por medio del lema de la matriz inversa, tomando la forma:

$$\boldsymbol{z}_\ell=\frac{\zeta_{\ell'}}{c_{\ell,1}}\Big(\boldsymbol{S}_{\ell'}-\boldsymbol{S}_{\ell'}\boldsymbol{A}_\ell^{-1}\hat{\boldsymbol{K}}_{\ell,\ell'}\Big)\boldsymbol{\alpha}_{\ell'} \tag{10.21}$$

El puntaje de una nueva muestra, $\boldsymbol{x}_*\in\mathbb{R}^P$, para el hiperplano de la clase ζ_ℓ es calculado como $f_\ell(\boldsymbol{x}_*)=\varphi(\boldsymbol{x}_*)^\top\boldsymbol{w}_\ell+b_\ell=\boldsymbol{s}_*^\top\boldsymbol{z}_\ell$, donde $\boldsymbol{s}_*=\begin{bmatrix}\varphi(\boldsymbol{x}_*)^\top & 1\end{bmatrix}^\top$. Luego, aplicando la función kernel, este se reescribe de la forma:

$$f_\ell(\boldsymbol{x}_*)=\frac{\zeta_{\ell'}}{c_{\ell,1}}\Big(\hat{\boldsymbol{k}}_{*,\ell'}^\top-\hat{\boldsymbol{k}}_{*,\ell}^\top\boldsymbol{A}_\ell^{-1}\hat{\boldsymbol{K}}_{\ell,\ell'}\Big)\boldsymbol{\alpha}_{\ell'}, \tag{10.22}$$

donde $\boldsymbol{k}_{*,\ell}\in\mathbb{R}^{N_\ell}$ contienen los elementos $\boldsymbol{k}_{*,\ell}=\{\kappa(\boldsymbol{x}_*,\boldsymbol{x}_n)\,:\,t_n=\zeta_\ell\}$, y $\hat{\boldsymbol{k}}_{*,\ell}=\boldsymbol{k}_{*,\ell}+\mathbf{1}_\ell$. Para encontrar la función de decisión, primero reescribimos $\boldsymbol{z}_\ell$ como:

$$\boldsymbol{z}_\ell=\begin{bmatrix}\boldsymbol{w}_\ell\\ b_\ell\end{bmatrix}=\frac{\zeta_{\ell'}}{c_{\ell,1}}\begin{bmatrix}\boldsymbol{\Phi}_{\ell'}\boldsymbol{\alpha}_{\ell'}+\boldsymbol{\Phi}_\ell\boldsymbol{A}_\ell^{-1}\hat{\boldsymbol{K}}_{\ell,\ell'}\boldsymbol{\alpha}_{\ell'}\\ \mathbf{1}_{\ell'}^\top\boldsymbol{\alpha}_{\ell'}+\mathbf{1}_\ell^\top\boldsymbol{A}_\ell^{-1}\hat{\boldsymbol{K}}_{\ell,\ell'}\boldsymbol{\alpha}_{\ell'}\end{bmatrix}, \tag{10.23}$$

entonces, la clasificación basada en ETWSVM no-lineal pude ser extraída de las ecuaciones (9.50), (10.22) y (10.23) como sigue:

$$t_* = \underset{\ell\in\{+,-\}}{\arg\min} \frac{|f_\ell(\boldsymbol{x}_*)|}{\|\boldsymbol{w}_\ell\|_{\mathcal{H}}}, \tag{10.24}$$

donde $\|\boldsymbol{w}_\ell\|_{\mathcal{H}}^2 = \langle \boldsymbol{w}_\ell, \boldsymbol{w}_\ell \rangle_{\mathcal{H}}$, y es calculado como:

$$\|\boldsymbol{w}_\ell\|_{\mathcal{H}}^2 = \frac{1}{c_{\ell,1}^2} \boldsymbol{\alpha}_{\ell'}^\top \Big(\boldsymbol{K}_{\ell',\ell'} - 2\boldsymbol{K}_{\ell',\ell}\boldsymbol{A}_\ell^{-1}\hat{\boldsymbol{K}}_{\ell,\ell'} + \hat{\boldsymbol{K}}_{\ell',\ell}\boldsymbol{A}_\ell^{-1}\boldsymbol{K}_{\ell,\ell}\boldsymbol{A}_\ell^{-1}\hat{\boldsymbol{K}}_{\ell,\ell'} \Big) \boldsymbol{\alpha}_{\ell'} \tag{10.25}$$

En resumen, la extensión no-lineal de nuestro enfoque ETWSVM explota los fundamentos de la representación kernel a través de un mapeo adecuado a un espacio de Hilbert con kernel reproductivo durante la optimización de los hiperplanos; mientras que ofrece una forma cerrada para el calculo del puntaje y las etiquetas de salida basadas en matrices kernel controladas por datos.

10.2. Aprendizaje de la función kernel en ETWSVM usando alineación centrada

Dentro del reconocimiento de patrones, en promedio, se prefiere utilizar el kernel Gaussiano por su capacidad de generalización, aproximación universal, y trazabilidad matemática. De ahí, que el siguiente función kernel surja en nuestro clasificador no-Lineal ETWSVM:

$$\kappa_{\boldsymbol{\Sigma}}(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) = \exp\left(-\frac{1}{2} (\boldsymbol{x}_n - \boldsymbol{x}_{n'})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{x}_{n'}) \right), \tag{10.26}$$

donde $\boldsymbol{\Sigma} \in \mathbb{R}^{P\times P}$ es la matriz de covarianza del kernel Gaussiano. Con el fin de evitar problemas de inestabilidad con respecto a la inversa de $\boldsymbol{\Sigma}$ y de encontrar estructuras de datos discriminativos, usamos una estrategia de alineación de kernel centralizado (*Centered Kernel Alignment*–(CKA)) para estimar $\boldsymbol{\Sigma}$. En particular, CKA aprovecha un ajuste controlado por datos de la matriz de covarianza del kernel Gaussiano al cuantificar la similitud entre la matriz $\boldsymbol{K}$ (con elementos dados por la ecuación (10.26)) y un kernel calculado sobre las etiquetas de salida $\boldsymbol{Z} \in \mathbb{R}^{N\times N}$. Este último, $z_{n,n'} = \delta(t_n - t_{n'})$, siendo $\delta(\cdot,\cdot)$ la función delta. Luego, reescribimos la inversa de la matriz de covarianza como: $\boldsymbol{\Sigma}^{-1} = \boldsymbol{E}\boldsymbol{E}^\top$, donde $\boldsymbol{E} \in \mathbb{R}^{P\times P'}$ ($P' \leqslant P$), y redefinimos la ecuación (10.26) explorando una proyección lineal, para que:

$$\kappa_{\boldsymbol{\Sigma}}(\boldsymbol{x}_n, \boldsymbol{x}_{n'}; \boldsymbol{E}) = \exp\left(-\frac{1}{2} \|\boldsymbol{x}_n\boldsymbol{E} - \boldsymbol{x}_{n'}\boldsymbol{E}\|_2^2 \right). \tag{10.27}$$

La función kernel en la ecuación (10.27) fomenta la alineación entre las matrices $\boldsymbol{K}$ y $\boldsymbol{Z}$ según la siguiente función de costo basada en CKA [60]:

$$\rho(\boldsymbol{K}(\boldsymbol{E}), \boldsymbol{Z}) = \left(\frac{\left\langle \widetilde{\boldsymbol{K}}(\boldsymbol{E}), \widetilde{\boldsymbol{Z}} \right\rangle_F}{\left\| \widetilde{\boldsymbol{K}}(\boldsymbol{E}) \right\|_F \left\| \widetilde{\boldsymbol{Z}} \right\|_F} \right), \tag{10.28}$$

donde $\widetilde{\boldsymbol{K}} \in \mathbb{R}^{N\times N}$ representa una matriz kernel centralizado calculado como $\widetilde{\boldsymbol{K}} = \widetilde{\boldsymbol{I}}\boldsymbol{K}\widetilde{\boldsymbol{I}}$, $\widetilde{\boldsymbol{I}} = \boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^\top/N$; donde $\langle\cdot,\cdot\rangle_F$ y $\|\cdot\|_F$ denotan el producto interno y la norma de Frobenius basada en matrices, respectivamente. Además, $\widetilde{\boldsymbol{K}}(\boldsymbol{E})$ destaca la dependencia del kernel Gaussiano sobre la proyección lineal $\boldsymbol{E}$. En efecto, la función de costo 10.28 es usado para aprender la inversa de la matriz de covarianza $\boldsymbol{\Sigma}^{-1}$ desde $\boldsymbol{E}$ como sigue [38]:

$$\boldsymbol{E}_* = \underset{\boldsymbol{E}}{\arg\max} \log(\rho(\boldsymbol{K}(\boldsymbol{E}), \boldsymbol{Z})), \tag{10.29}$$

donde la función $\log$ aumenta la convergencia del algoritmo. El problema de optimización en la ecuación (10.29) puede ser resuelta empleando el enfoque de gradiente descendiente [37]:

$$\begin{aligned}\nabla_{\boldsymbol{E}}(\rho(\boldsymbol{K}(\boldsymbol{E}),\boldsymbol{Z}))=&-4\boldsymbol{X}\left((\boldsymbol{G}\circ\boldsymbol{K}(\boldsymbol{E}))-\cdots\right.\\&\left.\cdots\operatorname{diag}\left(\mathbf{1}^\top(\boldsymbol{G}\circ\boldsymbol{K}(\boldsymbol{E}))\right)\right)\boldsymbol{X}^\top\boldsymbol{E},\end{aligned}\tag{10.30}$$

donde $\operatorname{diag}(\cdot)$ y el $\circ$ denota el operador diagonal y el producto de Hadamard, y $\boldsymbol{G}\in\mathbb{R}^{N\times N}$ es el gradiente de la función objetivo con respecto a $\boldsymbol{K}(\boldsymbol{E})$, calculada como:

$$\boldsymbol{G}=\frac{\widetilde{\boldsymbol{Z}}}{\operatorname{tr}\left(\boldsymbol{K}(\boldsymbol{E})\widetilde{\boldsymbol{Z}}\right)}-\frac{\widetilde{\boldsymbol{K}}(\boldsymbol{E})}{\operatorname{tr}\left(\boldsymbol{K}(\boldsymbol{E})\widetilde{\boldsymbol{K}}(\boldsymbol{E})\right)},\tag{10.31}$$

donde $\operatorname{tr}(\cdot)$ representa el operador traza. Notablemente, nuestra estrategia basada en CKA brinda una similitud basada en Gauss que favorece la separabilidad de los datos a través de la concordancia entre la entrada y el kernel basado en las etiquetas. Por otra parte, se evita la inversión de la matriz de covarianza debido a la relación $\boldsymbol{\Sigma}^{-1}=\boldsymbol{E}_*\boldsymbol{E}_*^\top$.

10.3. ETWSVM para problemas Multi-clase

Para los problemas de clasificación multi-clase, la base de datos desbalanceada contiene etiquetas de salidas $t_n\in\{1,2,\ldots,R\}$, donde $R\in\mathbb{N}$ es el número de clases en $\boldsymbol{X}$. Entonces, la matriz de muestras de la r-ésima clase $\boldsymbol{X}_r\in\mathbb{R}^{P\times N_r}$ es construida, donde $N=\sum_{r=1}^{R}N_r$ y $r\in\{1,2,\ldots,R\}$.

10.3.1. ETWSVM One-versus-Rest

El esquema *One-versus-Rest*–(OvR) es bien conocido y es uno de los más populares enfoques para problemas multiclase a partir de clasificadores binarios. Este sería acoplado con nuestro ETWSVM para la construcción de R hiperplanos no-paralelos y por ende R QPPs independientes, uno por cada clase [25]. Formalmente, ETWSVM-OvR encuentra R hiperplanos de separación $f_r(\boldsymbol{x})=\varphi\big(\boldsymbol{x}^\top\boldsymbol{w}_r\big)+b_r=0$, como se muestra en la figura 10.1(a), tal que:

$$\begin{aligned}\min_{\boldsymbol{w}_r,b_r,\boldsymbol{\xi}_{r'}}\ &\tfrac{c_{1,r}}{2}\left(\|\boldsymbol{w}_r\|_{\mathcal{H}}^2+b_r^2\right)+\tfrac{1}{2}\|\boldsymbol{\Phi}_r^\top\boldsymbol{w}_r+\mathbf{1}_r b_r\|_{\mathcal{H}}^2+c_{2,r'}\mathbf{1}_{r'}^\top\boldsymbol{\xi}_{r'}\\&\text{s.t. }\left(\boldsymbol{\Phi}_{r'}^\top\boldsymbol{w}_r+\mathbf{1}_{r'}b_r\right)\zeta_{r'}\geq\mathbf{1}_{r'}-\boldsymbol{\xi}_{r'},\\&\quad\boldsymbol{\xi}_{r'}\geq\mathbf{0},\ \forall r\in\{1,\ldots,R\};\end{aligned}\tag{10.32}$$

donde $\boldsymbol{\Phi}_r\in\mathbb{R}^{Q\times N_r}$ reúne las muestras mapeadas de la clase r mientras que $\boldsymbol{\Phi}_{r'}\in\mathbb{R}^{Q\times N_{r'}}$ las muestras mapeadas de las demás clases. Adicional, la formulación dual de el problema de optimización de la ecuación (10.32) es llevado acabo a través de lema de la matriz inversa para resolver R QPPs como en la ecuación (10.20), fijando $\boldsymbol{K}_{r,r'}=\boldsymbol{\Phi}_r^\top\boldsymbol{\Phi}_{r'}\in\mathbb{R}^{N_r\times(N-N_r)}$. Consecutivamente, la etiqueta de salida de una nueva muestra, $\boldsymbol{x}_*$, puede ser inferida como:

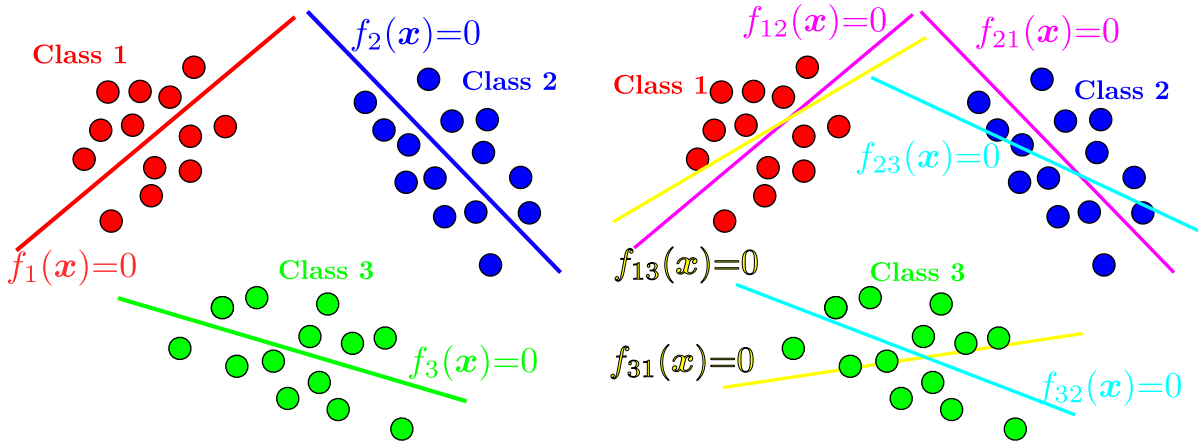
$$t_*=\underset{r\in\{1,\ldots,R\}}{\arg\min}\frac{|f_r(\boldsymbol{x}_*)|}{\|\boldsymbol{w}_r\|_{\mathcal{H}}},\tag{10.33}$$

donde $f_r(\boldsymbol{x}_*)$ y $\|\boldsymbol{w}_r\|_{\mathcal{H}}$ con calculados como en las ecuaciones (10.22) y (10.25), respectivamente.

10.3.2. ETWSVM One-versus-One

Por otra parte, el esquema *One-versus-One*–(OvO) acoplado con nuestro ETWSVM busca construir $R(R-1)/2$ clasificadores binarios [61], como se muestra en la figura 10.1(b). Explicitamente, para discriminar entre las clases r y r', ETWSVM-OvO almacena las matrices $\boldsymbol{\Phi}_r \in \mathbb{R}^{Q\times N_r}$ y $\boldsymbol{\Phi}_{r'} \in \mathbb{R}^{Q\times N_{r'}}$ para solucionar el QPPs de la ecuación (10.32) solamente para las dos clases de interés. En la etapa de decisión, ETWSVM-OvO adopta la estrategia de "voto mayoritario". Por ejemplo, si para una nueva muestra $\boldsymbol{x}_*$ el sub-clasificador entre las clases r-ésima y r'-ésima identifica la nueva etiqueta como r (evaluando la ecuación (10.33) únicamente para las dos clases bajo estudio), la votación para la r-ésima clase es aumentado en uno. Luego de evaluar todos los $R(R-1)/2$ sub-clasificadores, la etiqueta t_* es asignada por la clase con mayor número de votos [39].

Es importante mencionar que la estrategia de aprendizaje kernel presentada en la apartado 10.2 puede ser adoptado dentro de los ETWSVM-OvR y ETWSVM-OvO ya que los datos de entrada y los kernels basados en las etiquetas de salida ($\boldsymbol{K}$ y $\boldsymbol{Z}$) no dependen de la tarea en estudio, es decir, si es para clasificación binaria o multi-clase.



(a) One-versus-Rest (b) One-versus-One

Figura 10.1.: Representación geométrica de las estrategias multiclase acopladas con el ETWSVM. a) corresponde al esquema OvR acoplado con el ETWSVM, donde se generan R=3 hiperplano, uno por cada clase. b) representa el esquema OvO acoplado con el ETWSVM, donde para R=3, se general tres sub-clasificadores (por cada uno dos hiperplanos).

11. Montaje experimental

En este capitulo explicaremos las bases de datos usadas para probar nuestro ETWSVM, tanto para el problema de clasificación binaria como multi-clase. También el proceso de entrenamiento de nuestra propuesta y de los métodos con los cuales nos comparamos para demostrar las ventajas de la primera. Por último, las medidas de rendimiento utilizadas para evaluar todos los clasificadores.

11.1. Bases de Datos

Para ilustrar el rendimiento de nuestro ETWSVM, consideramos un par de base de datos sintéticas en dos dimensiones (P=2) devotas a clasificación binarias desbalanceadas: i) Unas media lunas (ver figura 12.1(a)) la cual posee una estructura no-lineal, fijando N_+=97 y N_-=227, y ii) dos grupos de datos generados a partir de dos distribuciones Gaussianas (ver figura 12.2), lo cual nos permite imponer diferentes tasas de desbalance ($\{1:2,\ldots,1:10\}$ fijando N_-=100) y escenarios de traslape (variamos la distancia Euclidea entre las medias de las clases desde el conjunto $[0, 10.14]$). Además, empleamos el públicamente disponible Repositorio de Aprendizaje de maquina UCI[1] para bases de datos de clasificación binaria como lo describen en [62, 63]. Con respecto al escenario multi-clase, también considerado el repositorio Keel[2] como es discutido en [25,36]. Las bases de datos contempla un amplio rango de campos, tales como patologías, vehículos, ingeniería, informática biológica, financiera, etc., de diferentes tamaños y tasas de desbalance. Una pequeña descripción de las bases de datos de los repositorios UCI y Keel se describe en el cuadro 11.1.

Binary classsification (UCI repository)

Name	IR	P	N	**Minority class**
BankNote	0.8005	4	1372	1
Wisconsin	0.5938	30	569	Malignant
Ionosphere	0.5600	34	351	b
Pima-Indians	0.5360	8	768	1
Biodeg	0.5093	41	1055	RB
Iris	0.5000	4	150	Iris Virginica
Haberman	0.3600	4	306	2
Transfusion	0.3123	4	748	1
Vehicle	0.3076	18	846	Van
cmc	0.2921	9	1473	2:Long-term
Housing	0.0743	13	506	1 (CHAS)
Balance	0.0851	4	625	Balanced

Multi-class classsification (Keel repository)

Name	N	P	R
Balance	625	4	3
Contraceptive	1473	9	3
Dermatology	358	34	6
Ecoli	336	7	8
Glass	214	7	7
Hayes-roth	160	4	3
New-thyroid	215	5	3
Iris	150	4	3
Thyroid	7200	21	3
Wine	178	13	3
Penbased	10992	16	10

Tabla 11.1.: Repositorios UCI y Keel descritos para clasificación desbalanceada. $IR = N_+/N_-$: tasa de desbalance , P: # de características, N: # de muestras, y R: # de clases.

[1] https://archive.ics.uci.edu/ml/index.php
[2] https://sci2s.ugr.es/keel/category.php?cat=clas

11.2. Entrenamiento de ETWSVM y métodos de comparación

11.2.1. Pre-procesamiento

Considerando la tarea de clasificación binaria, un *ν-nearest neighbor*–(ν-NN),método de sub-muestreo, es llevado a cabo antes del aprendizaje del kernel de ETWSVM para eludir la influencia de los valores atípicos y para mitigar la carga computacional cuando inferimos la proyección lineal en la ecuación (10.27). Por consiguiente, un grafo local es definido para revelar la densidad intraclase en el grupo mayoritario siguiendo una filosofía familiar como en [30]. Sea $\boldsymbol{U}\in\mathbb{R}^{N_-\times N_-}$ una matriz adyacente definida con elementos:

$$u_{nn'} = \begin{cases} 1 & \boldsymbol{x}_n\in\Omega_\nu(\boldsymbol{x}_{n'}) \text{ y } \boldsymbol{x}_{n'}\in\Omega_\nu(\boldsymbol{x}_n) \\ 0 & \text{Otro caso,} \end{cases} \tag{11.1}$$

donde Ω_ν contiene los $\nu\in\mathbb{N}$ vecinos más cercanos de $\boldsymbol{x}_n$ en $\boldsymbol{X}_-$ acorde a la distancia Euclidea ($\forall\boldsymbol{x}_n, \boldsymbol{x}_{n'}\in\boldsymbol{X}_-$). Luego, es calculado el vector, $\boldsymbol{\vartheta}\in\mathbb{R}^{N_-}$, de coeficientes de sub-muestreo como: $\vartheta_n=\sum_{\boldsymbol{x}_{n'}\boldsymbol{X}_-} u_{nn'}$. La matriz de la clase mayoritaria puede ser reducida por el número de muestras de la clase minoritaria, N_+ con los valores más altos del coeficiente de sub-muestreo en $\boldsymbol{\vartheta}$. En consecuencia, las muestras provenientes de regiones de alta densidad son más propensos ha ser seleccionados que las muestras provenientes de regiones de baja densidad, por ejemplo, se descartan los valores atípicos. Del mismo modo, para la tarea multi-clase, la tarea del sub-muestreo es llevada a cabo para eliminar datos de entrada forzando los cada uno de los $R-1$ clases (excluyendo la clase con más bajo número de muestras) a contener el siguiente número de instancias: $\hat{N}=\min(\min_{r'} RN_r, N_r)$. El factor R alienta la preservación de la estructura de los datos de entrada respecto al número de clases proporcionadas. Para concretar la prueba, nosotros experimentalmente fijamos $\nu=8$ inspirado en una grilla de búsqueda discutida por los autores en [30].

11.2.2. Aprendizaje de la función kernel

La función kernel es obtenida a través del enfoque basado en CKA discutido en el apartado 10.2, configurando los siguientes parámetros de optimización: La inicialización de la proyección lineal dentro del iterativo gradiente descendiente es ajustada acorde al bien conocido enfoque *Principal Component Analysis*–(PCA), la tolerancia del gradiente descendiente es fijado en 10^{-6}, el máximo número de iteraciones es limitado empíricamente a 300, y el número de dimensiones relevantes (P') es ajustada para retener el $95\,\%$ de la varianza explicada [38]. En aras de simplicidad, consideramos ETWSVM* como el emparejamiento entre el preprocesamiento basado en ν-NN (apartado 11.2.1), el aprendizaje de la función kernel basado en CKA (apartado 11.2.2), y la discriminación basada en el ETWSVM (ver figura 11.1). Asimismo, ETWSVM*-OvO y ETWSVM*-OvR sostienen un multi-clase extensión de ETWSVM*.

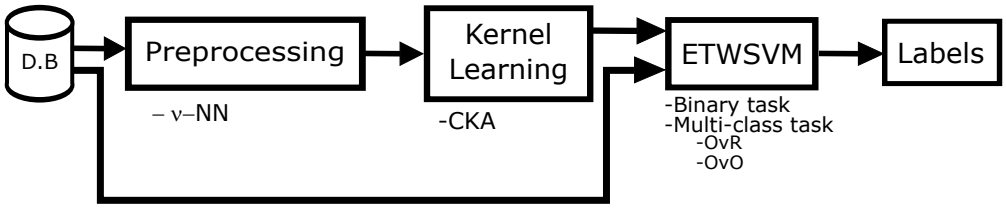


Figura 11.1.: Disposición del ETWSVM*. Un submuestreo basado en ν-NN es aplicado antes de la etapa del aprendizaje de la función kernel. Entonces, la clasificación binaria o multiclase es realizado para resolver el problema de optimización de ETWSVM.

11.2.3. Evaluación y sintonización de ETWSVM

Para la sintonización, de los parámetros libres de los clasificadores, y la evaluación de los mismos, tanto binarios como multiclase, usamos un esquema de validación cruzada anidada 10-*folds*. Para el escenario binario, usamos las medidas de rendimiento *Accuracy*–(Acc), la *Geometric Mean*–(GM), y la *F-measure*–(FM), las cuales son calculadas como sigue:

$$\begin{aligned} Acc =& \frac{T_P + T_N}{N_{test}} \\ GM =& \sqrt{Sen \times Spec} \\ FM =& \frac{2Pre \times Rec}{Pre + Rec}, \end{aligned} \tag{11.2}$$

donde $T_P \in \mathbb{N}$ es el número de muestras positivas que son estimadas como positivas, $T_N \in \mathbb{N}$ es el número de muestras negativas que son estimadas como negativas, y $N_{test} \in \mathbb{N}$es el número de instancias en el conjunto de prueba. $Sen{=}T_P/(T_P + F_N)$ y $Spec{=}T_N/(T_N + F_P)$ son las medidas de sensibilidad y especificidad, respectivamente; donde $F_N \in \mathbb{N}$ es el número de muestras positivas que son clasificadas como negativas y $F_P \in \mathbb{N}$ es el número de muestras negativas que son clasificadas como positivas. Por otra parte, $Pre{=}T_P/(T_P + F_P)$ y $Rec = Sen$ son los valores de las medidas de precisión y recuperación, respectivamente. En general, el Acc es la medida de rendimiento, en clasificación, más popular; sin embargo, esta no es indicada para calcular el rendimiento en problemas de clasificación desbalanceada, un ejemplo es supongamos una base de datos la cual tenga una tasa de desbalance es de $9:1$, si el clasificador sesga por completo a la clase minoritaria y clasifica bien todas las muestras provenientes de la clase mayoritaria su $Acc{=}90\,\%$. En consecuencia, la evaluación basada en GM y FM son también adoptadas, ya que tienen menos probabilidades de sufrir el problema de desbalance, dado que tienen en cuenta la distribución de clase [8].

Adicional, para evaluar el escenario multi-clase, contemplamos las medidas de rendimiento *Mean Accuracy*–($\overline{Acc}$) y *F-measure*$_\mu$–(FM_μ). El primero captura la efectividad promedio del clasificador junto con las R clases mientras el último favorece la capacidad predictiva considerando las etiquetas positivas [64]. Estas medidas son calculadas como sigue:

$$\begin{aligned} \overline{Acc} =& \frac{1}{R}\sum_{r=1}^{R}\frac{T_{P_r} + T_{N_r}}{N_{test}} \\ FM_\mu =& \frac{2}{R}\frac{\overline{Pre} \times \overline{Rec}}{\overline{Pre} + \overline{Rec}}, \end{aligned} \tag{11.3}$$

donde $\overline{Pre}{=}\frac{1}{R}\sum_{r=1}^{R} Pre_r$ y $\overline{Rec}{=}\frac{1}{R}\sum_{r=1}^{R} Rec_r$. El subíndice r indica que la cantidad correspondiente se computará en la clase r-ésima.

Los parámetros de regularización son buscados del conjunto $\{2^{-7}, 2^{-5}, \ldots, 2^5, 2^7\}$ con respecto a la evaluación basada en FM. No obstante, para la reducción de la complejidad computacional, en nuestros experimentos, fijamos $c_{1,\ell}{=}c_{1,\ell'}$ y $c_{2,\ell}{=}c_{2,\ell'}$. Nótese que el ETWSVM sencillo incluye un kernel Gaussiano con covarianza isótropa $\boldsymbol{\Sigma}{=}\sigma^2\boldsymbol{I}$, donde el valor del ancho bando $\sigma^2 \in \mathbb{R}^+$ es buscado dentro del rango $\{0.1\sigma_0, 0.2\sigma_0, \ldots, \sigma_0\}$ con respecto al valor de FM, $\sigma_0{=}\text{med}(\|\boldsymbol{x}_n - \boldsymbol{x}_{n'}\|_2)$ y med$(\cdot)$ indica el operador mediana (no se aplica la etapa de preprocesamiento). Todos los QPPs relacionados con nuestro ETWSVM-son resueltos utilizando el algoritmo de punto-interior-convexo del Toolbox de Optimización de MATLAB, fijando la tolerancia de el método en 10^{-12}, la terminación en el primer modo óptimo y las tolerancias de

violación de las restricciones en 10^{-8}, y el máximo número de iteraciones en 200.

11.2.4. Métodos de comparación

Junto al ETWSVM sencillo (el cual incluye un kernel Gaussiano con covarianza isótropa) y el ETWSVM* (el cual contiene el preprocesamiento basado en ν-NN y el aprendizaje de la función kernel basado en CKA), también contemplamos el bien conocido método de sobre-muestreo SMOTE en nuestra propuesta. Entonces, nos referimos a ETWSVM** como la implementación del algoritmo SMOTE para mejorar la tasa de desbalance; entonces, la base de datos con sobre-muestreo es usada para el aprendizaje de kernel basado en CKA y la construcción de la frontera de decisión. El número de vecinos más cercanos dentro del SMOTE es fijado desde el conjunto $\{1, 3, 6, 9, 12\}$ como es discutido en [18]. Aquí, SMOTE es únicamente usado con clasficadores binarios, debido a su creciente costo computacional [7, 65]. Sin embargo, la evaluación del ETWSVM** en problemas binarios nos dará información sobre las ventajas y desventajas del algoritmo SMOTE en comparación al ETWSVM y ETWSVM*. La implementación en MATLAB de nuestros ETWSVM, ETWSVM* y ETWSVM** (incluyendo ambas opciones, binaria y multi-clase) están públicamente disponibles[3].

Para realizar una buena comparación, los siguientes métodos del estado del arte dedicados a la clasificación binaria son examinados: i) SVM [66], ii) WSVM [67], iii)TBSVM [59], iV) WLTSVM [30], y V) SVM$_{SMOTE}$ [68]. El toolboxes de estadistica y aprendizaje de máquina de MATLAB son usados para implementar los clasificadores SVM y WSVM. Los scripts de WLTSVM son posteados por sus autores en [30][4]. Asimismo, los algoritmos TBSVM y SMOTE son implementados como se menciona en [18, 59]. La versión no-lineal prueba ser más competitiva que la alternativa lineal para todos los métodos proporcionados en el estado del arte, y garantiza una comparación justa con nuestros algoritmos basados en ETWSVM, fijando un kernel Gaussiano isótropa. El ancho de banda, σ, es buscado como en el ETWSVM sencillo. Adicional, los siguientes enfoques multi-clase son estudiados: SVM-OvR, SVM-OvO, TBSVM-OvR, y TBSVM-OvO [39]. De nuevo, una función kernel Gaussiana isótropa es utilizada. Todos los experimentos fueron realizados en a entorno de MATLAB 9.1 en una computadora con un procesador i5 de 4^{ta} generación (2.0GHz) y 6GB de memoria RAM.

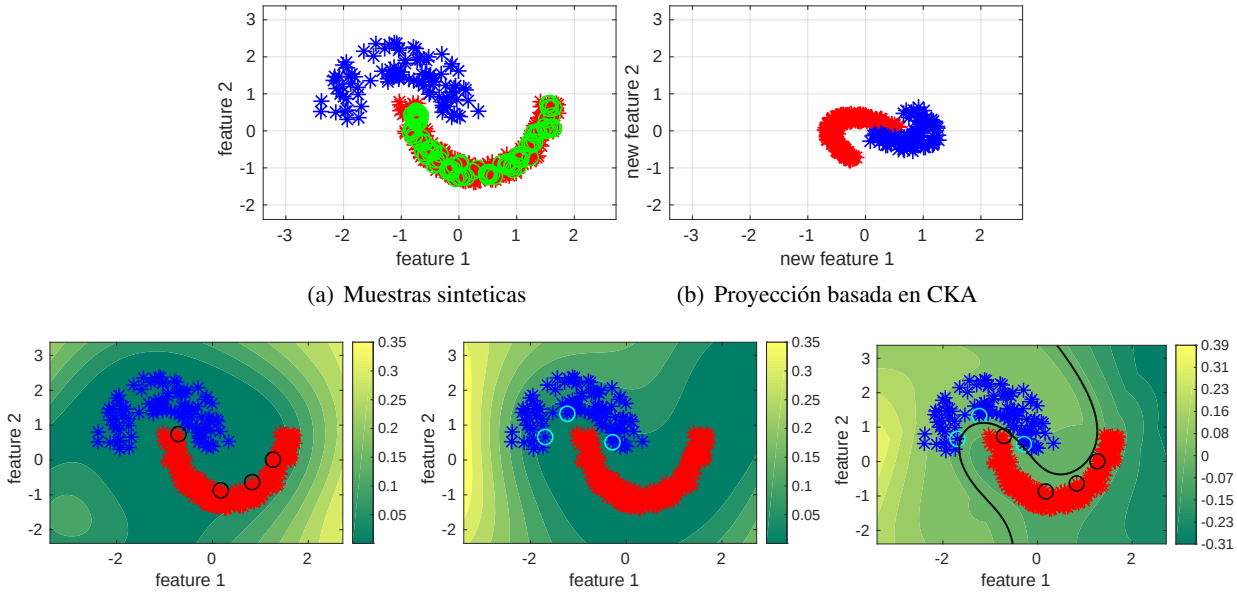
[3] https://github.com/cralji/ETWSVM_PR.git
[4] http://www.optimal-group.org/Resource/LTSVM.html

Parte III.

Resultados, Conclusiones y Trabajos Futuros

12. Resultados y Discusión

12.1. Resultados en datos sintéticos



(a) Muestras sinteticas (b) Proyección basada en CKA

(c) Distancia al hiperplano minoritario (d) Distancia al hiperplano mayoritario (e) Frontera de decisión

Figura 12.1.: Resultados de la clasificación del ETWSVM* sobre la base de datos media-luna. (a) muestra los datos sintéticos de media-luna, las muestras azules provienen de la clase minoritaria mientras las muestras rojas de la clase mayoritaria; las instancias encerradas en círculos verdes son seleccionadas por el submuestreo basado en ν-NN. (b) muestra la proyección basada en CKA fijando P'=2. (c) y (d) muestra la distancia entre los datos de entrada y los hiperplanos minoritario y mayoritaria, respectivamente; donde las muestras encerradas en negro y en cían representa los vectores de soporte de los hiperplanos minoritario y mayoritario, respectivamente. (e) muestra la frontera de decisión y la mínima distancia entre (c) y (d) (ver función de decisión del ETWSVM en ecuación (10.24)).

Primero, un ejemplo representativo es llevado a cabo sobre la base de datos de media-luna para inspeccionar visualmente el rendimiento del ETWSVM*. Inicialmente, nosotros sólo describimos los resultados del ETWSVM* para dilucidar las virtudes de la representación respecto al mapeo no-lineal para un sostenible RKHS. En la figura 12.1(a) se muestra los datos de entrada y las muestras seleccionadas de la clase mayoritaria a través de la etapa del preprocesamiento basado en el ν-NN. Como se puede ver la etapa del ν-NN captura las muestras relevantes codificando la estructura de luna. Junto a, la proyección basada en

CKA, en este experimento lo fijamos P'=2 para visualizar el propósito, las metas para encontrar un compromiso entre "estiramiento" y "aplastamiento" para revelar información relevante (ver figura 12.1(b)). El espacio proyectado, el cual codifica dentro de la Gaussiana multi-variada (ver ecuación (10.27)), compensando la diferencia entre las densidades de las clases mayoritaria y minoritaria colapsando las instancias adecuadamente. Notablemente, las "nuevas nubes de puntos" concentrando la estructura fundamental de cada grupo. Por lo tanto, la distancia a los hiperplanos minoritarios y mayoritarios son mostrados en las figuras 12.1(c) y 12.1(d) exhibiendo como ETWSVM* codificando el funcionamiento de la media-luna. En efecto, ETWSVM* sólo requiere siete vectores de soporte ($\boldsymbol{\alpha} > 0$) para construir una frontera suave como se ve en la figura 12.1(e).

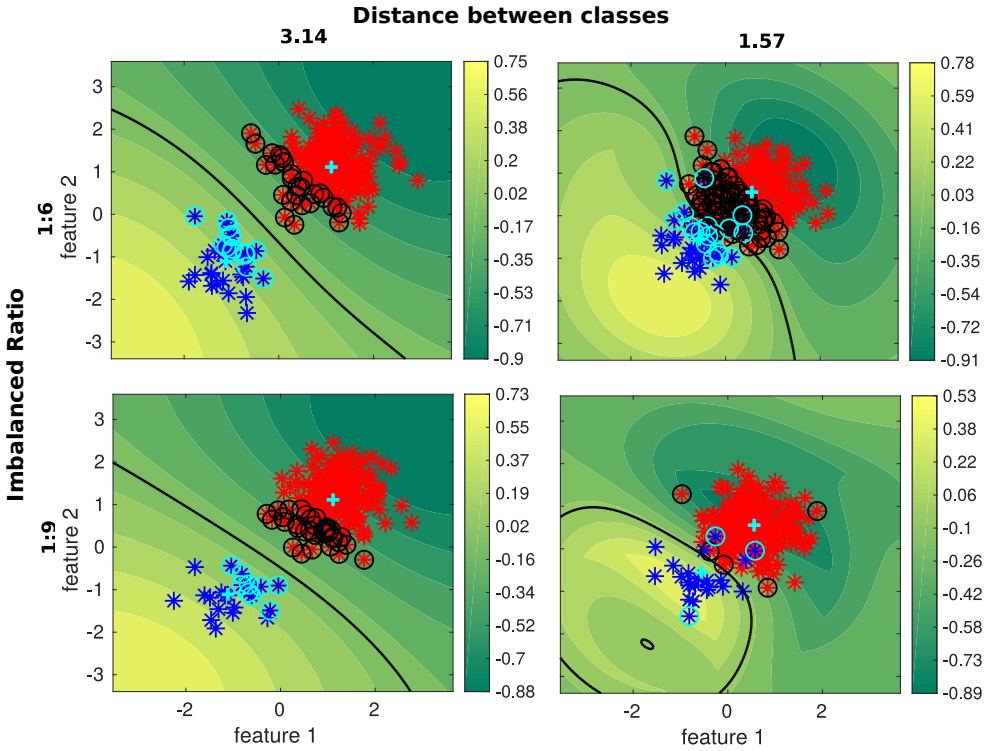


Figura 12.2.: Inspección visual de los resultados sobre bases de datos sintéticas generada a partir de dos distribuciones Gaussianas multivariadas (clasificación basada en ETWSVM*). Diferentes escenarios de desbalance y traslape entre las clases son probadas. Las muestras azules provienen de la clase minoritaria mientras que las rojas pertenecen a la mayoritaria. Las muestras enceradas en negro y cían representan los vectores de soporte de las clase minoritaria y mayoritaria, respectivamente, y la cruz la media de las distribuciones. La frontera de decisión es representada con una linea negra y la mínima distancia a los hiperplanos es coloreada en el fondo.

Segundo, la figura 12.2 presenta las fronteras de decisión del ETWSVM* para las bases de datos generadas a partir de dos Gaussianas multivariadas considerando varios escenarios de desbalance y traslape entre las clases. Como se puede apreciar, una tasa alta de desbalance influencia a la frontera de decisión en estar más cerrada a la clase minoritaria; también, un alto traslape entre clases, causa un incremento en los vectores de soporte de los hiperplanos. A demás, las figuras 12.3 y 12.4 muestran el porcentaje de GM y FM resultado de la clasificación sobre bases de datos generadas a partir de dos distribuciones Gaussianas en función de la tasa de desbalance, el traslape de las clases, y los valores de los parámetros de regularización. En efecto, tales parámetros fueron variados desde el conjunto $c_{1,\ell}, c_{2,\ell} \in \{10^{-4}, 10^{-1}, 1, 10^{1}, 10^{2}\}$, y fijamos la función

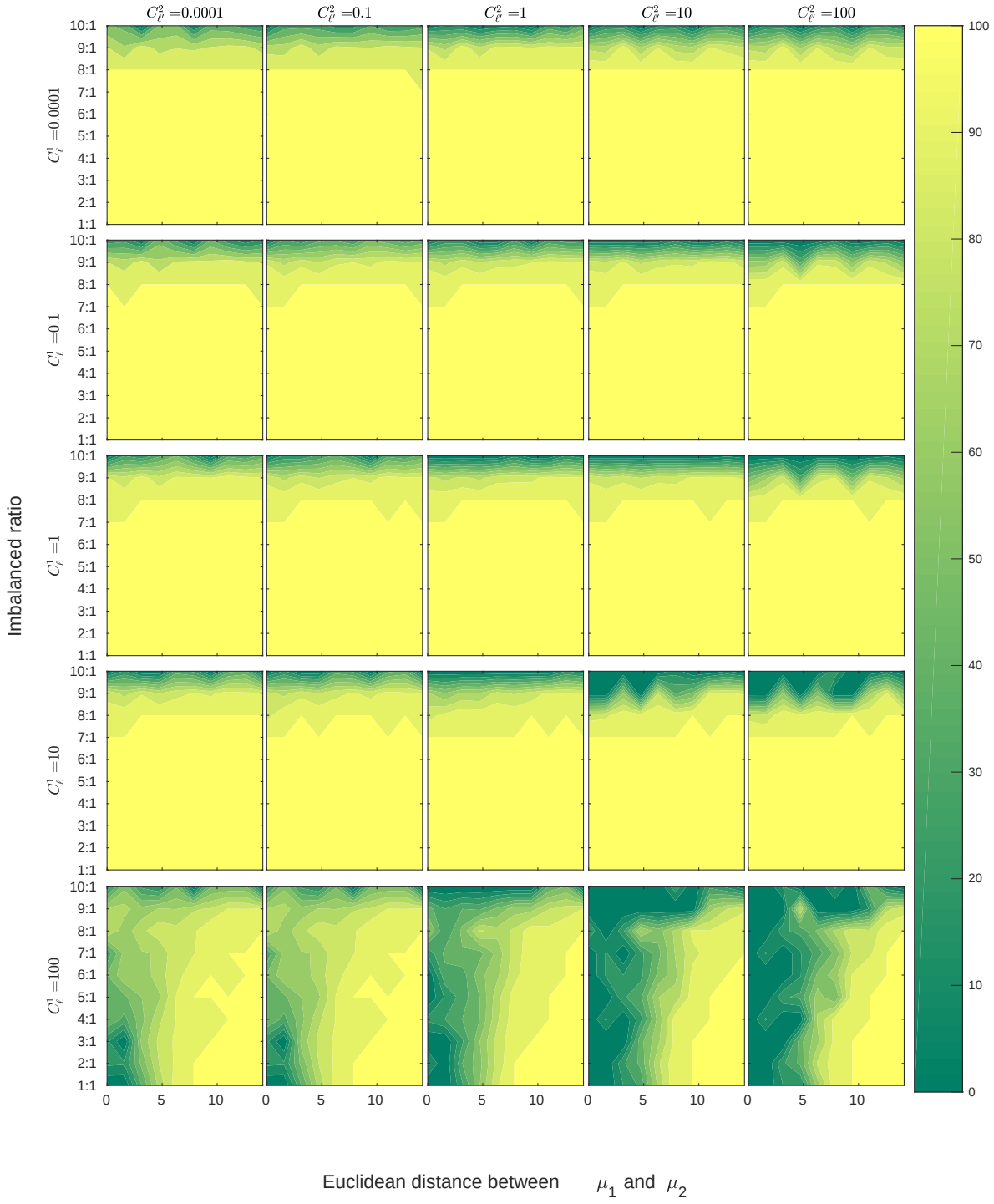


Figura 12.3.: Análisis de los parámetros de regularización (resultados de clasificación del ETWSVM*). La evaluación cualitativa basada en el GM se representa variando la tasa de desbalance y la distancia entre las medias de cada clase. Dentro de cada gráfica un par de valores de $c_{1,\ell}$ y $c_{2,\ell\prime}$ es fijado para solucionar los QPPs en la ecuación (10.20). Esta medida de rendimiento se calcula variando de abajo hacia arriba la relación de desbalance y de izquierda a derecha la distancia entre las clases.

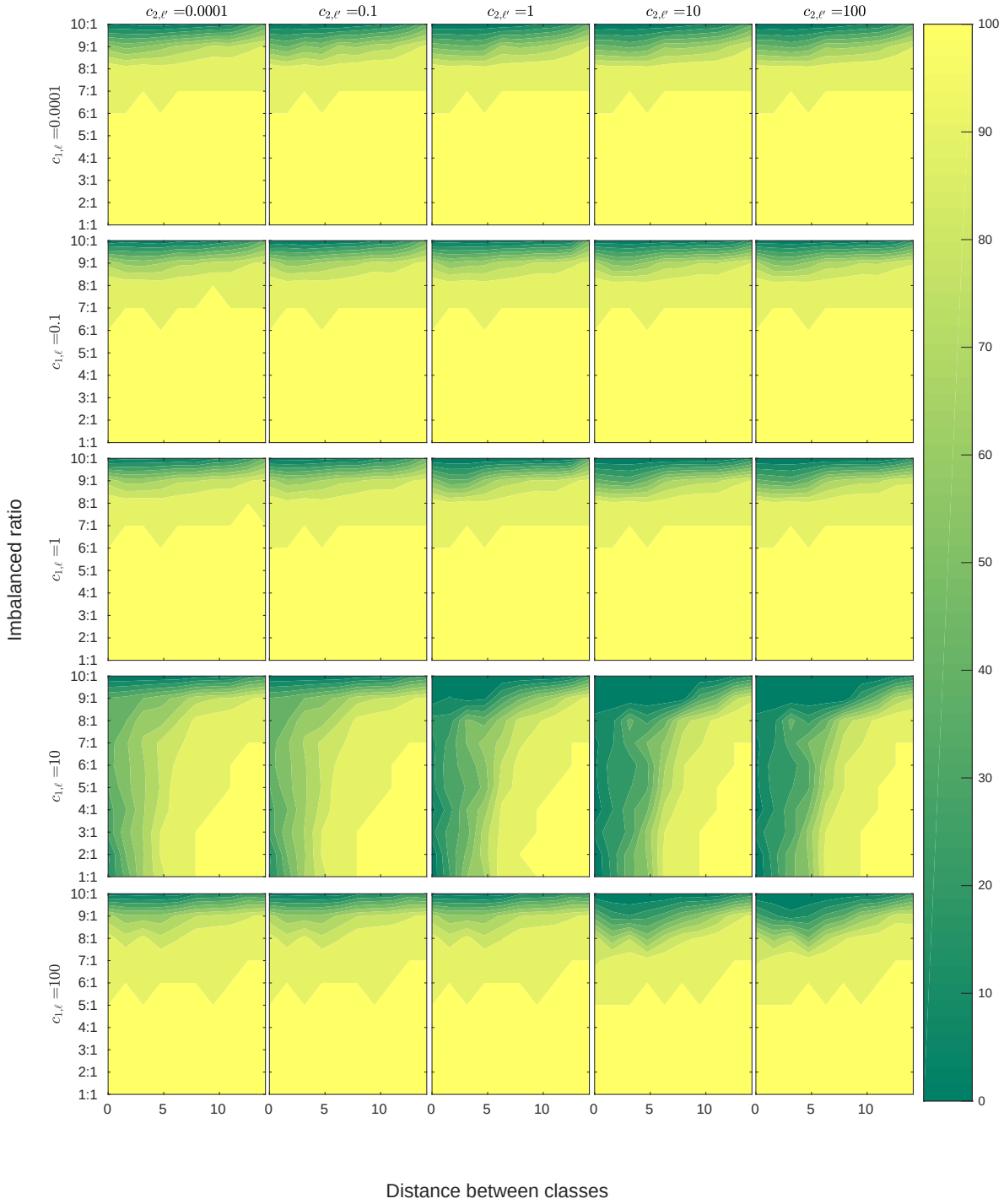


Figura 12.4.: Análisis de los parámetros de regularización (resultados de clasificación del ETWSVM*). La evaluación cualitativa basada en el FM se representa variando la tasa de desbalance y la distancia entre las medias de cada clase. Dentro de cada gráfica un par de valores de $c_{1,\ell}$ y $c_{2,\ell\prime}$ es fijado para solucionar los QPPs en la ecuación (10.20). Esta medida de rendimiento se calcula variando de abajo hacia arriba la relación de desbalance y de izquierda a derecha la distancia entre las clases.

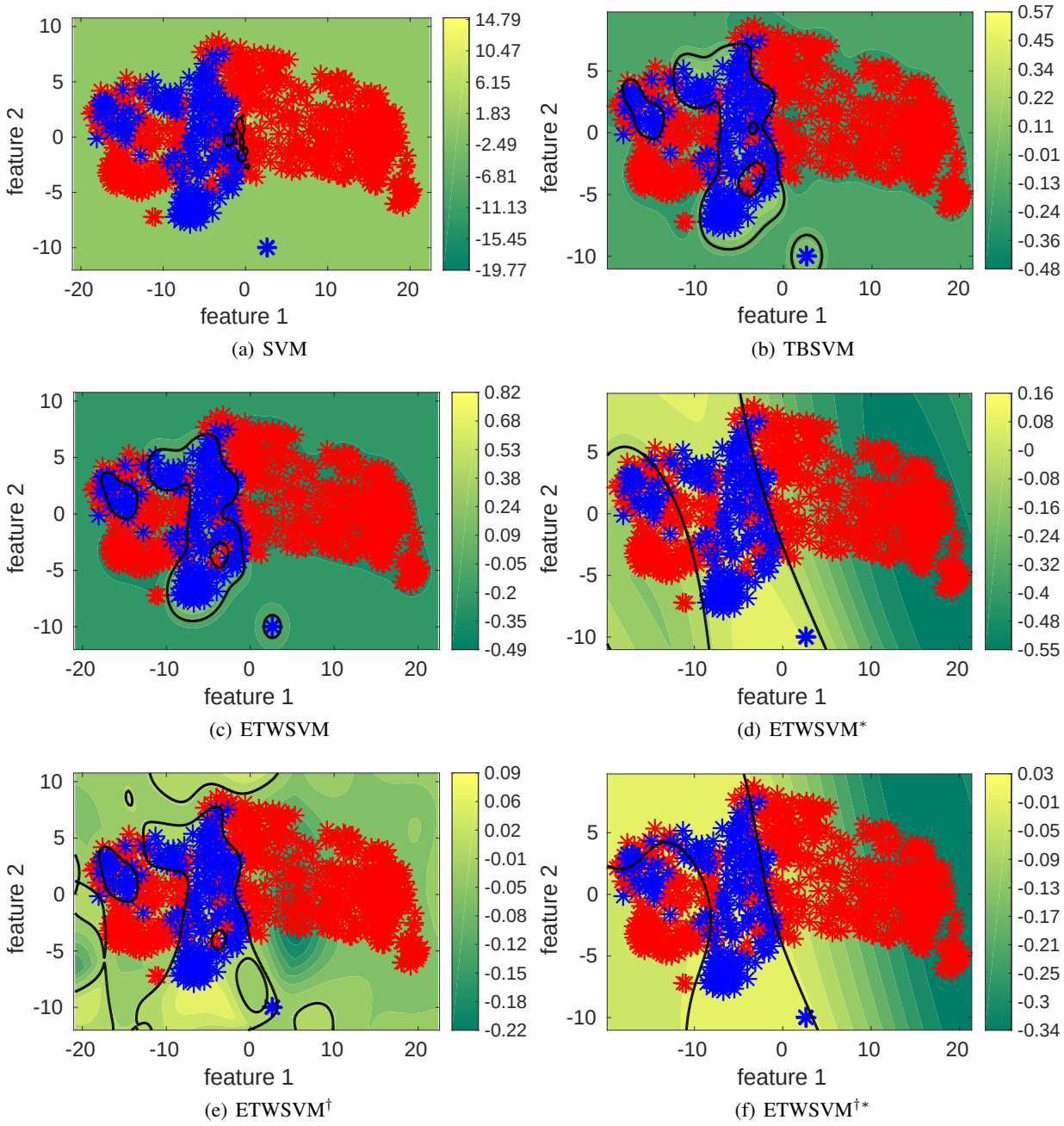
kernel a través de la proyección basada en CKA. Cómo es claro, el parámetro $c_{1,\ell}$ es más sensible a la tasa de desbalance, y el clasificador refleja mejores rendimientos para valores bajos de $c_{1,\ell}$. Lo anterior puede ser explicado por la influencia de tal parámetro en la matriz inversa de la matriz Hessiana (ver ecuación (10.19)). Por el contrario, el parámetro $c_{2,\ell'}$ permite compensar el traslape de las clases por la penalización de la mala-clasificación de las muestras provenientes de la clase $\zeta_{\ell'}$ (ver ecuación (10.2)). De ahí, después de la inspección visual en los resultados del FM en la figura 12.4 podemos fijar $c_{1,\ell}{=}10^{-3}$ y $c_{2,\ell'}{=}1$ para obtener una compensación entre el rendimiento del clasificador y el costo computacional llevado acabo por la sintonización de estos parámetros, cediendo a las siguientes extensiones del ETWSVM: ETWSVM†, ETWSVM$^{\dagger *}$ y ETWSVM$^{\dagger **}$ († presenta los valores de los parámetros de regularización fijados).

12.2. Resultados en bases de datos reales: clasificación binaria

Para ilustrar lo propuesto, aplicamos el bien conocido algoritmo *t-student stochastic neighbor embedding (t-SNE)* para incrustar la base de datos Vehicle dentro de un espacio de características de dos dimensiones (2D) [69]. Aquí, nuestro objetivo es visualizar la capacidad de generalización de nuestro algoritmo respecto a la compleja frontera de decisión en el problema de clasificación en el mundo real. La figura 12.5 muestra la frontera de decisión de: SVM, TBSVM, ETWSVM, y ETWSVM$*$, ETWSVM†, y ETWSVM$^{\dagger *}$ en el espacio 2D embebido por el t-SNE. Como se puede observar, el SVM no es capaz de codificar la estructura fundamental de los datos y esta completamente sesgado hacia la clase mayoritaria. El TBSVM construye una frontera de decisión que encierra la clase minoritaria, lo cual puede ser producto de sobre-aprendizaje sobre esta clase. Por otra parte, el ETWSVM y ETWSVM†, los cuales incluyen un mapeo adecuado a RKHS, pueden codificar las estructuras de la clase mayoritaria y minoritaria pero a expensas de una frontera de decisión compleja. Adicional, el ETWSVM* y ETWSVM$^{\dagger *}$, los cuales incluyen el aprendizaje del kernel por medio de una proyección lineal basada en CKA, puede revelar la información discriminatoria manteniendo una frontera de decisión suave que garantiza una clasificación confiable con capacidad de generalización.

Sucesivamente, las tablas 12.1 y 12.2 presentan el rendimiento de clasificación y tiempo de entrenamiento para las bases de datos del repositorio UCI. Notamos que nuestro ETWSVM y sus variaciones (ETWSVM*, ETWSVM**, ETWSVM†, ETWSVM$^{\dagger *}$ y ETWSVM$^{\dagger **}$); muestran resultados aceptables en comparación a los clasificadores SVM y WSVM sobre las bases de datos Haberman, Housing, Vehicle, Transfusion, Ionosphere, Balance, Biodeg, e Iris. Para el resto de bases de datos nuestros clasificadores superan a SVM y WSVM. Con respecto al algoritmo SVM$_{SMOTE}$, los clasificadores basados en ETWSVM son comparables con respecto al desempeño de la clasificación para todas las bases de datos estudiadas. Sin embargo, la clasificación por medio del SVM$_{SMOTE}$ requiere un alto costo computacional debido a su etapa de sobre-muestreo (ver la tabla 12.2). En particular, nuestros métodos basados en ETWSVM pueden lidiar con la desafiante tasa de desbalance, incluso en ausencia de una etapa de re-muestreo para la clasificación. Además, los resultados del WLTSVM son comparables con nuestras propuestas en las bases de datos Haberman, Prima-Indians, BankNote e Iris. No obstante, para el resto de problemas de clasificación binaria, nuestros clasificadores exceden al rendimiento del WLTSVM. Ahora, para el algoritmo TBSVM, se pueden observar altos valores de Acc, pero tiende a obtener resultados sesgados, por ejemplo, viendo el desempeño con respecto al GM y el FM sobre las bases de datos Housing, Transfusion, y Balance.

En promedio, nuestros planteamientos logran los más altos valores de Acc, GM, y FM en promedio, y ganan en 10 de 12 bases de datos para todas la medidas de rendimiento calculadas. Notablemente, el ETWSVM*, el cual incluye el preprocesamiento basado en ν-NN y el aprendizaje de la función kernel por medio de la proyección lineal basada en CKA, consigue los siguientes resultados en promedio: $86\,\%$ Acc, $79\,\%$ GM, y $69\,\%$, los cuales son estadísticamente similares a los resultados de ETWSVM y ETWSVM**. Además,



(a) SVM (b) TBSVM

(c) ETWSVM (d) ETWSVM*

(e) ETWSVM† (f) ETWSVM†*

Figura 12.5.: Frontera de decisión de la base de datos Vehicle después de la embebimiento a un espacio 2D a través del t-SNE. Las muestras azules provienen de la clase minoritaria mientras que las rojas pertenecen a la clase mayoritaria. Las fronteras de decisión son mostradas en negro, y el puntaje de la clasificación (o la mínima distancia entre a los hiperplanos) son coloreados en el fondo.

después establecimos los valores de los parámetros de regularización en ETWSVM* como se explica en el apartado 12.1, ETWSVM†*, el desempeño muestra una ligera disminución (87 % Acc, 72 % GM, y 64 % FM) pero se obtiene una ganancia significativa con respecto al tiempo computacional (ver tabla 12.2). De

	SVM	$WSVM$	SVM_{SMOTE}	$WLTSVM$	$TBSVM$	$ETWSVM$	$ETWSVM^{*}$	$ETWSVM^{**}$	$ETWSVM^{\dagger}$	$ETWSVM^{\dagger *}$	$ETWSVM^{\dagger **}$
	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc
Dataset	GM	GM	GM	GM	GM	GM	GM	GM	GM	GM	GM
	FM	FM	FM	FM	FM	FM	FM	FM	FM	FM	FM
	88, 3±3, 7	92, 3±13, 0	97, 5±1, 9	91, 8±1, 4	91, 9±2, 2	97, 5±1, 4	96, 7±1, 7	96, 2±1, 7	**97, 6±1, 1**	96, 7±1, 4	96, 4±1, 3
BankNote	88, 7±3, 6	86, 6±30, 5	97, 4±2, 1	92, 2±1, 3	91, 9±2, 1	97, 5±1, 4	96, 7±1, 7	96, 3±1, 6	**97, 6±1, 1**	96, 8±1, 3	96, 6±1, 2
	87, 8±3, 6	86, 4±30, 4	97, 1±2, 2	91, 3±1, 4	43, 9±30, 4	97, 2±1, 6	96, 3±1, 9	95, 8±1, 8	**97, 3±1, 3**	96, 4±1, 4	96, 1±1, 4
	51, 3±5, 8	97, 2±1, 5	97, 2±2, 1	80, 8±5, 6	93, 0±2, 9	**98, 2±2, 2**	97, 7±2, 0	98, 1±1, 9	97, 0±2, 8	96, 7±3, 4	97, 4±2, 4
Wisconsin	46, 6±9, 2	97, 0±1, 7	96, 8±2, 3	81, 5±5, 5	90, 4±4, 1	**97, 9±2, 3**	97, 2±2, 5	97, 7±2, 5	96, 4±2, 9	96, 1±4, 3	97, 0±2, 8
	60, 6±3, 0	96, 2±2, 0	96, 2±2, 8	77, 0±6, 1	54, 3±0, 5	**97, 6±2, 9**	96, 9±2, 8	97, 4±2, 7	96, 0±3, 7	95, 3±4, 8	96, 4±3, 3
	94, 6±2, 5	94, 0±2, 5	93, 7±5, 0	82, 4±6, 8	92, 3±5, 9	**94, 6±5, 0**	93, 7±5, 7	93, 7±3, 7	94, 0±3, 2	88, 7±5, 0	88, 3±5, 3
Inosphere	**93, 6±3, 1**	93, 1±3, 3	92, 2±6, 7	82, 5±7, 6	93, 2±5, 5	93, 5±6, 1	91, 8±8, 4	92, 4±4, 5	92, 7±4, 0	84, 3±8, 2	83, 5±7, 7
	92, 3±3, 5	91, 5±3, 6	90, 8±7, 8	78, 3±8, 2	51, 3±5, 6	92, 2±7, 4	90, 4±9, 7	90, 9±5, 4	91, 4±4, 8	82, 0±9, 5	81, 1±9, 3
	25, 3±4, 7	25, 0±4, 3	72, 8±5, 0	65, 4±7, 0	74, 4±3, 6	73, 6±3, 7	**75, 5±3, 5**	74, 0±5, 4	72, 5±5, 1	71, 5±2, 9	73, 8±4, 8
Prima-Indians	24, 5±5, 6	24, 5±5, 0	72, 7±4, 6	68, 3±7, 1	71, 6±4, 1	72, 3±3, 7	**73, 9±3, 1**	73, 3±5, 4	72, 7±5, 7	65, 3±5, 2	69, 0±5, 9
	30, 3±6, 1	30, 3±5, 0	65, 4±5, 2	64, 7±5, 5	51, 7±0, 4	64, 8±4, 4	66, 6±3, 7	**65, 9±6, 2**	65, 3±6, 4	56, 1±6, 3	60, 7±7, 6
	87, 0±2, 9	87, 3±3, 0	87, 8±2, 2	77, 8±2, 2	84, 1±5, 3	87, 4±1, 5	86, 6±1, 4	88, 1±2, 0	**88, 7±1, 7**	86, 4±4, 3	**86, 8±3, 6**
Biodeg	86, 6±2, 8	86, 5±3, 0	**86, 8±2, 7**	75, 2±2, 2	84, 3±4, 6	85, 6±2, 7	84, 4±2, 6	86, 5±2, 6	86, 8±2, 6	85, 0±4, 8	85, 2±3, 9
	81, 7±3, 7	81, 8±3, 9	82, 3±3, 3	67, 7±2, 6	54, 8±7, 2	81, 2±2, 6	79, 8±2, 7	82, 3±2, 9	**83, 0±2, 9**	80, 2±6, 2	80, 6±5, 1
	86, 0±9, 1	99, 3±2, 1	99, 3±2, 1	99, 3±2, 1	99, 3±2, 1	99, 3±2, 1	**100, 0±0, 0**	**100, 0±0, 0**	99, 3±2, 1	**100, 0±0, 0**	**100, 0±0, 0**
Iris	73, 9±19, 3	98, 9±3, 3	98, 9±3, 3	98, 9±3, 3	98, 9±3, 3	98, 9±3, 3	**100, 0±0, 0**	**100, 0±0, 0**	98, 9±3, 3	**100, 0±0, 0**	**100, 0±0, 0**
	69, 8±23, 7	98, 9±3, 5	98, 9±3, 5	98, 9±3, 5	50, 0±0, 0	98, 9±3, 5	**100, 0±0, 0**	**100, 0±0, 0**	98, 9±3, 5	**100, 0±0, 0**	**100, 0±0, 0**
	62, 8±8, 5	71, 5±5, 0	61, 8±5, 2	59, 5±16, 3	**73, 9±3, 9**	73, 2±6, 2	68, 3±7, 8	70, 6±9, 0	61, 7±14, 8	71, 9±5, 3	69, 3±8, 3
Haberman	53, 4±12, 1	54, 5±9, 5	56, 2±12, 5	38, 3±22, 9	32, 0±24, 9	**63, 8±10, 0**	58, 5±7, 8	57, 0±14, 0	62, 5±13, 0	41, 0±10, 1	53, 5±10, 7
	37, 5±12, 9	**50, 1±11, 4**	40, 7±13, 5	28, 1±18, 1	41, 9±1, 2	50, 1±12, 0	44, 0±10, 0	42, 6±16, 2	49, 4±13, 8	26, 5±11, 1	38, 7±11, 7
	69, 1±4, 1	68, 0±5, 0	62, 6±10, 3	36, 0±31, 2	23, 8±0, 4	76, 0±5, 7	76, 2±4, 6	73, 0±7, 2	75, 4±6, 2	**79, 4±3, 4**	74, 2±4, 9
Transfusion	64, 8±5, 9	59, 7±22, 0	51, 9±28, 2	39, 2±33, 9	0, 0±0, 0	64, 6±7, 4	65, 7±7, 3	**68, 3±7, 1**	66, 1±6, 7	54, 7±8, 8	66, 4±8, 3
	47, 4±6, 2	44, 0±17, 0	39, 2±21, 6	49, 0±2, 9	38, 4±0, 6	50, 0±9, 6	51, 0±8, 8	**52, 7±8, 5**	51, 2±9, 6	42, 4±11, 1	50, 5±9, 3
	97, 5±1, 4	97, 9±1, 3	97, 6±1, 5	82, 4±4, 6	85, 7±7, 0	97, 9±2, 1	95, 2±2, 9	96, 1±2, 4	**98, 5±1, 4**	96, 3±1, 3	96, 5±2, 4
Vehicle	97, 7±2, 0	97, 7±1, 8	96, 9±2, 4	82, 2±4, 0	64, 8±44, 8	96, 8±3, 0	94, 0±5, 5	96, 5±2, 9	**97, 9±2, 1**	95, 3±2, 7	96, 4±2, 7
	94, 9±3, 0	95, 6±2, 7	95, 0±3, 1	69, 1±6, 4	38, 1±0, 5	95, 5±4, 2	89, 9±6, 5	92, 3±4, 7	**96, 7±2, 9**	92, 3±2, 8	92, 8±4, 8
	26, 8±4, 7	33, 2±3, 2	67, 8±6, 2	31, 9±22, 4	64, 6±6, 3	68, 0±4, 1	66, 2±4, 2	65, 9±3, 3	67, 6±3, 5	**75, 6±2, 7**	68, 4±4, 0
cmc	19, 9±17, 5	31, 7±4, 1	58, 8±21, 1	31, 7±22, 6	66, 0±5, 2	66, 6±3, 4	65, 2±3, 5	65, 6±3, 3	66, 8±5, 4	41, 5±7, 5	**67, 1±5, 4**
	21, 8±18, 9	44, 0±4, 3	41, 9±15, 5	28, 1±7, 3	36, 8±0, 6	47, 9±3, 5	46, 1±4, 0	46, 5±3, 7	47, 8±5, 9	26, 0±8, 0	**48, 3±6, 0**
	91, 9±2, 7	88, 6±5, 3	90, 3±3, 7	44, 9±6, 6	**93, 1±1, 0**	89, 7±2, 7	88, 4±5, 9	83, 4±3, 8	89, 2±4, 1	91, 5±2, 3	86, 0±4, 0
Housing	63, 4±16, 7	67, 7±27, 6	66, 3±25, 1	46, 2±18, 2	0, 0±0, 0	47, 8±29, 0	41, 6±38, 4	**72, 3±11, 5**	43, 7±39, 6	38, 7±27, 3	62, 0±26, 0
	42, 0±14, 9	41, 1±19, 4	**42, 8±21, 2**	13, 7±3, 7	12, 9±1, 7	27, 8±19, 9	27, 3±32, 3	34, 2±10, 2	26, 8±25, 9	24, 7±18, 1	31, 8±16, 3
	90, 7±3, 3	90, 7±3, 2	83, 8±6, 0	49, 5±20, 7	92, 2±0, 5	83, 2±5, 9	84, 3±4, 7	83, 2±6, 7	84, 2±5, 0	**93, 4±1, 8**	85, 8±4, 6
Balance	81, 1±18, 9	57, 3±11, 2	69, 3±13, 4	52, 5±16, 5	0, 0±0, 0	75, 8±13, 3	78, 4±13, 0	69, 8±21, 0	78, 7±17, 2	63, 5±13, 0	**82, 2±10, 8**
	54, 7±19, 4	**59, 0±13, 2**	37, 2±13, 6	18, 5±8, 8	11, 0±5, 3	40, 6±15, 5	42, 9±12, 7	36, 9±16, 5	43, 3±15, 8	49, 5±14, 1	47, 1±12, 5
	72, 6±25, 8	78, 8±25, 2	84, 4±14, 4	66, 8±22, 3	80, 7±20, 6	86, 6±11, 4	85, 7±11, 7	85, 2±12, 0	85, 5±13, 1	**87, 3±10, 3**	85, 2±11, 4
Averange	66, 2±25, 9	71, 3±26, 0	78, 7±18, 0	65, 7±23, 1	57, 8±39, 2	80, 1±17, 3	79, 0±18, 4	**81, 3±15, 1**	80, 1±17, 9	71, 9±23, 9	79, 9±15, 9
	60, 1±25, 2	68, 2±25, 8	69, 0±26, 8	57, 0±28, 9	40, 4±14, 8	70, 3±26, 2	69, 3±25, 9	69, 8±25, 9	**70, 6±26, 0**	64, 3±29, 9	68, 7±25, 1
	0/12	0/12	0/12	0/12	2/12	2/12	2/12	1/12	3/12	4/12	2/12
Wins	1/12	0/12	1/12	0/12	0/12	4/12	2/12	3/12	2/12	1/12	3/12
	1/12	2/12	1/12	0/12	0/12	2/12	1/12	3/12	3/12	1/12	2/12

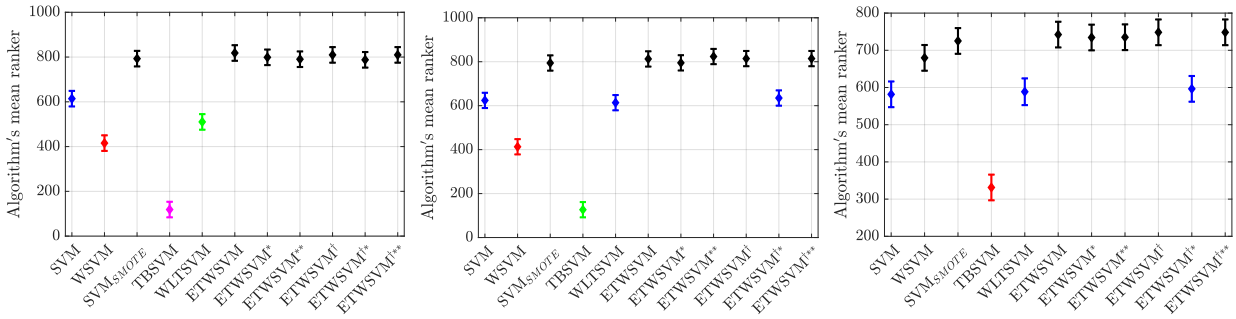
Tabla 12.1.: Resultados sobre las bases de datos del repositorio UCI (clasificación binaria). Las medidas Acc, GM y FM son consideradas. La media ± la desviación estándar son mostradas como resultado de la validación cruzada anidada de 10-fold.

Datasets	SVM	WSVM	SVM_{SMOTE}	WLTSVM	TBSVM	EWTSVM	$EWTSVM^{*}$	$EWTSVM^{**}$	$EWTSVM^{\dagger}$	$EWTSVM^{\dagger *}$	$EWTSVM^{\dagger **}$
BankNote	7, 7±0, 3	**7, 4±0, 7**	59, 6±3, 2	102, 9±5, 2	217, 7±5, 5	215, 3±27, 9	127, 1±18, 7	135, 3±2, 5	170, 1±2, 4	49, 8±1, 2	92, 0±1, 1
Wisconsin	2, 1±0, 1	**1, 8±0, 0**	17, 6±0, 3	13, 0±0, 6	25, 0±0, 5	29, 9±0, 7	13, 6±0, 6	15, 0±0, 2	31, 2±0, 7	4, 3±0, 3	7, 7±0, 2
Ionosphere	1, 4±0, 1	**1, 1±0, 0**	11, 3±0, 3	5, 7±0, 3	10, 9±0, 1	14, 5±0, 3	5, 3±0, 0	8, 1±0, 2	14, 9±0, 1	1, 7±0, 1	3, 1±0, 1
Prima-Indians	12, 2±0, 5	12, 8±0, 6	128, 8±5, 2	22, 9±0, 9	52, 7±1, 0	57, 8±6, 2	17, 5±0, 2	25, 3±0, 2	50, 1±0, 7	**7, 2±0, 3**	13, 2±0, 4
Biodeg	9, 0±0, 3	**8, 4±0, 5**	89, 2±1, 5	47, 8±2, 2	119, 3±1, 8	104, 0±1, 0	39, 4±5, 0	58, 6±6, 9	104, 6±1, 1	13, 0±0, 3	24, 5±0, 5
Iris	0, 5±0, 1	**0, 4±0, 0**	3, 7±0, 0	1, 7±0, 0	6, 6±0, 1	5, 7±0, 0	1, 8±0, 1	2, 5±0, 1	6, 1±0, 1	0, 4±0, 0	1, 4±0, 1
HaberMan	6, 1±0, 6	3, 9±0, 3	49, 7±2, 6	4, 6±0, 6	10, 6±0, 4	17, 9±1, 4	3, 1±0, 2	5, 0±0, 4	12, 5±0, 1	**0, 9±0, 2**	3, 1±0, 1
Transfusion	14, 2±3, 8	13, 7±1, 7	79, 8±17, 0	20, 4±0, 7	53, 2±0, 6	52, 9±0, 7	11, 0±0, 8	33, 2±0, 8	53, 0±0, 8	**2, 5±0, 1**	35, 5±1, 0
Vehicle	3, 9±0, 1	3, 6±0, 1	43, 6±0, 6	27, 4±1, 0	72, 5±2, 3	68, 8±1, 0	14, 6±0, 4	48, 7±1, 8	69, 7±1, 0	**3, 5±0, 2**	50, 7±1, 5
cmc	22, 5±6, 3	26, 3±7, 5	377, 8±74, 9	92, 9±5, 2	317, 0±3, 6	256, 9±16, 4	54, 9±2, 2	203, 2±13, 7	249, 2±3, 3	**11, 8±0, 5**	189, 4±4, 3
Housing	2, 1±0, 1	2, 0±0, 1	42, 3±1, 8	8, 2±0, 5	27, 6±0, 6	30, 7±0, 9	3, 7±0, 1	20, 4±1, 2	31, 4±0, 8	**0, 4±0, 0**	25, 3±0, 8
Balance	8, 6±0, 8	10, 4±1, 8	149, 8±5, 9	15, 2±0, 7	38, 3±0, 4	42, 4±0, 9	5, 5±0, 1	31, 8±2, 0	42, 6±0, 8	**0, 5±0, 0**	40, 5±1, 5
Averange	**7, 5±6, 4**	7, 6±7, 5	87, 8±101, 7	30, 2±34, 0	79, 3±95, 7	74, 7±80, 5	24, 8±36, 0	48, 9±60, 4	69, 6±72, 7	8, 0±13, 9	40, 5±53, 6
Win	0/12	5/12	0/12	0/12	0/12	0/12	0/12	0/12	0/12	7/12	0/12

Tabla 12.2.: Tiempo de entrenamiento para las bases de datos del repositorio UCI (clasificación binaria). La media ± desviación estándar son representan el tiempo de entrenamiento en segundos por cada fold.

ahí, podemos expresar que el $ETWSVM^{\dagger *}$ es la mejor opción con respecto a la relación al desempeño de la clasificación y el tiempo de entrenamiento.

También, se realizo una prueba estadística entre los clasificadores, realizamos la prueba de Kruskal-Wallis para comparar las medidas de desempeño entre los métodos. Si la hipótesis nula para la igualdad de medianas es rechazada, realizamos múltiples pruebas de comparación usando el algoritmo Tukey-Kramer para estudiar la diferencia entre los clasificadores [70, 71]. Todos los niveles de significancia con medidos en $5\,\%$. Como se puede ver en la figura 12.6 SVM. WSVM, TBSVM, y WLTSVM exhiben comportamientos inestables a lo largo de las medidas Acc, GM y FM. Por otra parte, los algoritmos basados en ETWSVM y SVM$_{SMOTE}$ son estadísticamente iguales. Aunque, el ETWSVM$^{\dagger *}$ muestra diferencias estadísticas en GM y FM en comparación a los demás algoritmos basados en ETWSVM y el SVM$_{SMOTE}$ debido a la relación entre el rendimiento de la clasificación y el tiempo de entrenamiento, aún así se mantiene como una alternativa adecuada para lidiar con lo desafíos del desbalance y traslape de los dato sin necesidad de a conocimiento a priori en relación con el ajuste del algoritmo.



(a) Acc (b) GM (c) FM

Figura 12.6.: Prueba de diferencia estadística entre los clasificadores basada en Tukey-Kramer. Se muestra el diagrama de caja del ranking de media del algoritmo, y el color indica que algoritmos son estadísticamente iguales.

12.3. Resultados bases de datos reales: clasificación multiclase

En la tabla 12.3 muestra el desempeño de la clasificación multiclase. Por un lado, SVM-OvR y SVM-OvO presentan resultados similares ($\overline{Acc} \sim 88\,\%$ y FM$_{\mu} \sim 78\,\%$). Por el otro lado, el TBSVM-OvO supera al algoritmo TBSVM-OvR; de hecho, el TBSVM-OvR presenta el más bajo desempeño entre los clasificadores estudiados. Esto último puede explicarse porque la estrategía OvR tiende a incrementar los problemas relacionados al desbalance de datos. Respecto a las variantes del ETWSVM, podemos ver como nuestras propuestas contrarresta, a través de un adecuado mapeo a un RKHS, los problemas de desbalance de datos independientemente de la estrategia multiclase (OvO o OvR). En general, el ETWSVM supera los métodos SVM y TBSVM, demostrando su capacidad de generalización de nuevo. En contraste con el problema de clasificación binario, el ETWSVM$^{\dagger *}$ no presenta el más bajo costo computacional por la necesidad de la estimación del kernel basado en CKA en varias ocasiones debido a los esquemas OvR y OvO.

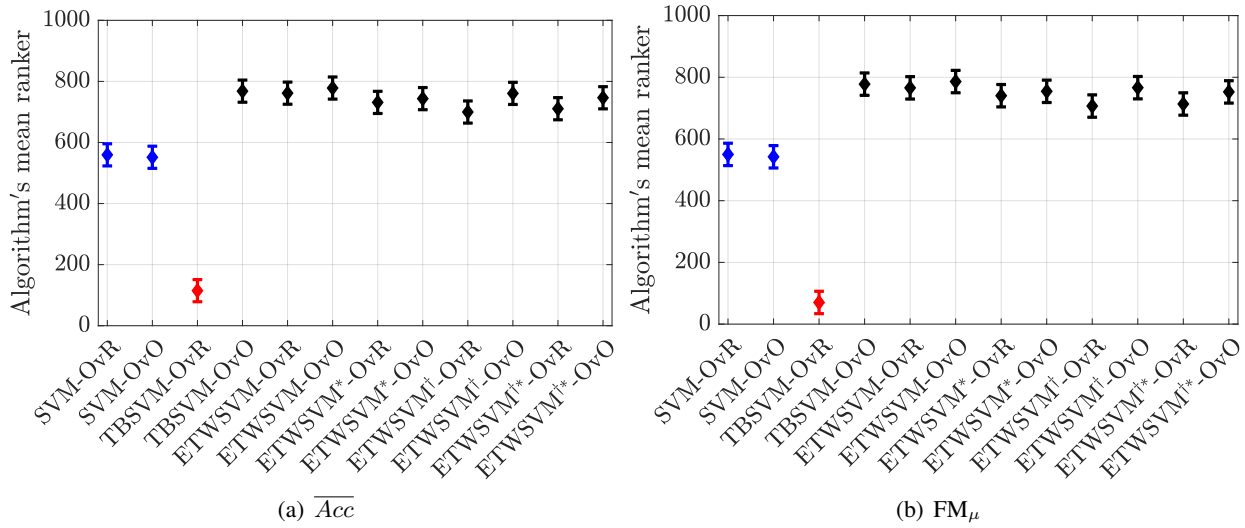
Por último, realizamos las mismas pruebas estadísticas del escenario biclase. Como es evidente en la figura 12.7, los clasificadores basados en ETWSVM y el TBSVM-OvO presentan resultados consistentes en $\overline{Acc}$ y FM$_{\mu}$, mientras que, SVM-OvO, SVM-OvR, y TBSVM-OvR son estadísticamente diferentes de acuerdo a la prueba realizada.

Datasets	SVM OvR	SVM OvO	TBSVM OvR	TBSVM OvO	ETWSVM OvR	ETWSVM OvO	ETWSVM* OvR	ETWSVM* OvO	ETWSVM† OvR	ETWSVM† OvO	ETWSVM†* OvR	ETWSVM†* OvO
	$\overline{Acc}$	$\overline{Acc}$	$\overline{Acc}$	$\overline{Acc}$	$\overline{Acc}$	$\overline{Acc}$	$\overline{Acc}$	$\overline{Acc}$	$\overline{Acc}$	$\overline{Acc}$	$\overline{Acc}$	$\overline{Acc}$
	FM_μ	FM_μ	FM_μ	FM_μ	FM_μ	FM_μ	FM_μ	FM_μ	FM_μ	FM_μ	FM_μ	FM_μ
Balance	94, 4±2, 3	94, 4±2, 3	38, 2±0, 9	99, 0±1, 2	98, 6±1, 2	98, 4±1, 3	98, 8±1, 2	98, 9±1, 0	98, 3±1, 5	98, 6±0, 9	97, 2±2, 4	**99, 1±1, 3**
	91, 7±3, 4	91, 7±3, 4	7, 4±1, 3	98, 6±1, 8	97, 9±1, 9	97, 6±1, 9	98, 2±1, 8	98, 4±1, 5	97, 4±2, 3	97, 9±1, 3	95, 8±3, 6	**98, 7±2, 0**
Contraceptive	61, 8±0, 1	61, 8±0, 1	56, 7±0, 9	70, 0±2, 2	**70, 3±2, 7**	70, 2±2, 0	70, 0±2, 1	68, 7±3, 8	70, 1±3, 1	70, 1±2, 7	68, 8±2, 1	68, 3±3, 2
	42, 7±0, 2	42, 7±0, 2	35, 1±1, 3	54, 9±3, 3	**55, 4±4, 1**	55, 3±3, 1	55, 1±3, 2	53, 0±5, 7	55, 2±4, 6	55, 2±4, 1	53, 2±3, 1	52, 5±4, 8
Dermatology	98, 6±0, 9	97, 5±1, 7	69, 6±1, 1	99, 1±0, 9	98, 6±0, 9	99, 1±0, 6	98, 8±1, 0	98, 9±0, 7	98, 7±0, 8	**99, 3±0, 7**	98, 2±0, 9	98, 7±0, 9
	95, 9±2, 6	92, 6±5, 1	8, 8±3, 4	97, 3±2, 6	95, 9±2, 6	97, 3±1, 8	96, 4±2, 9	96, 7±2, 2	96, 2±2, 4	**97, 8±2, 1**	94, 5±2, 6	96, 2±2, 7
Ecoli	95, 3±1, 8	95, 2±1, 8	80, 3±2, 4	96, 0±0, 9	96, 8±1, 2	96, 7±1, 0	96, 5±1, 0	96, 5±1, 5	**97, 0±1, 4**	96, 9±1, 2	96, 5±0, 9	96, 6±1, 4
	79, 0±7, 9	78, 3±7, 9	11, 5±10, 6	81, 9±3, 8	85, 8±5, 4	85, 1±4, 4	84, 3±4, 3	84, 0±6, 7	**86, 7±6, 1**	85, 8±5, 4	84, 3±4, 1	84, 6±6, 3
Glass	84, 9±2, 6	84, 4±2, 3	77, 3±1, 2	89, 2±3, 9	89, 1±3, 1	90, 0±4, 4	88, 2±3, 7	90, 1±2, 4	88, 8±3, 2	89, 7±3, 3	89, 0±4, 5	**90, 6±3, 0**
	54, 6±7, 6	53, 2±6, 9	31, 9±3, 7	67, 5±11, 7	67, 3±9, 4	70, 1±13, 1	64, 5±11, 1	70, 2±7, 1	66, 4±9, 5	69, 1±9, 9	67, 0±13, 4	**71, 9±9, 0**
Hayes-roth	70, 2±5, 3	69, 7±5, 1	63, 2±5, 7	88, 9±7, 8	87, 9±6, 3	87, 8±6, 9	86, 5±7, 9	86, 9±4, 5	**89, 4±7, 4**	88, 0±4, 6	84, 4±7, 5	87, 8±5, 6
	55, 3±7, 9	54, 6±7, 6	44, 7±8, 5	83, 3±11, 7	81, 9±9, 5	81, 7±10, 3	79, 7±11, 8	80, 4±6, 8	**84, 1±11, 1**	81, 9±6, 9	76, 7±11, 2	81, 7±8, 3
New-thyroid	97, 5±2, 5	97, 8±2, 1	57, 2±6, 9	96, 9±2, 5	97, 2±2, 8	96, 9±3, 0	90, 1±4, 6	97, 8±2, 6	97, 2±2, 2	97, 2±2, 3	**98, 1±2, 2**	97, 8±3, 0
	96, 3±3, 7	96, 7±3, 1	35, 8±10, 4	95, 4±3, 8	95, 8±4, 2	95, 3±4, 4	85, 1±6, 8	96, 7±3, 8	95, 8±3, 4	95, 8±3, 4	**97, 2±3, 3**	96, 7±4, 5
Iris	81, 3±9, 8	81, 3±7, 5	55, 6±0, 0	96, 4±4, 1	97, 3±2, 3	**97, 8±3, 1**	93, 8±6, 4	94, 7±2, 8	96, 9±2, 2	**97, 8±3, 1**	93, 3±6, 0	94, 2±5, 2
	72, 0±14, 7	72, 0±11, 2	33, 3±0, 0	94, 7±6, 1	96, 0±3, 4	**96, 7±4, 7**	90, 7±9, 5	92, 0±4, 2	95, 3±3, 2	**96, 7±4, 7**	90, 0±9, 0	91, 3±7, 7
Thyroid	95, 7±0, 9	95, 7±1, 1	43, 6±12, 2	**97, 7±0, 9**	97, 0±1, 5	96, 8±1, 3	95, 3±0, 7	97, 4±1, 5	96, 9±1, 7	97, 1±0, 9	97, 0±1, 6	97, 5±1, 4
	93, 5±1, 3	93, 6±1, 6	15, 4±18, 3	**96, 5±1, 3**	95, 6±2, 3	95, 2±2, 0	92, 9±1, 0	96, 1±2, 3	95, 3±2, 5	95, 7±1, 4	95, 6±2, 4	96, 3±2, 1
Wine	97, 0±2, 9	96, 3±3, 0	51, 3±1, 3	98, 9±1, 8	98, 5±2, 6	98, 9±1, 8	98, 9±1, 8	97, 4±3, 9	**99, 63±1, 2**	99, 2±2, 5	98, 5±1, 9	97, 4±2, 5
	95, 5±4, 4	94, 5±4, 5	27, 0±1, 9	98, 3±2, 7	97, 8±3, 9	98, 3±2, 7	98, 4±2, 6	96, 1±5, 9	**99, 44±1, 8**	98, 8±3, 7	97, 8±2, 9	96, 1±3, 8
Penbased	97, 6±0, 5	97, 6±0, 6	81, 9±0, 1	99, 6±0, 3	99, 6±0, 2	**99, 7±0, 2**	99, 4±0, 2	99, 6±0, 3	99, 6±0, 3	99, 7±0, 3	99, 6±0, 2	99, 7±0, 2
	87, 8±2, 7	87, 9±3, 2	9, 6±0, 6	97, 9±1, 5	97, 9±1, 0	**98, 5±1, 2**	96, 7±1, 1	98, 1±1, 3	97, 8±1, 5	98, 5±1, 5	97, 9±1, 1	98, 4±1, 0
Averange	88, 6±12, 6	88, 4±12, 6	61, 4±14, 6	93, 8±8, 8	93, 7±8, 7	93, 8±8, 7	92, 4±8, 7	93, 4±9, 1	93, 9±8, 7	**94, 0±8, 8**	92, 8±9, 2	93, 4±9, 1
	78, 6±19, 6	78, 0±19, 5	23, 7±13, 4	87, 8±14, 7	87, 9±14, 4	88, 3±14, 2	85, 6±14, 3	87, 4±14, 6	88, 2±14, 5	**88, 5±14, 5**	86, 4±14, 9	87, 7±14, 4
Wins	0/11	0/11	0/11	1/11	1/11	2/11	0/11	0/11	3/11	2/11	1/11	2/11
	0/11	0/11	0/11	1/11	1/11	2/11	0/11	0/11	3/11	2/11	1/11	2/11

Tabla 12.3.: Resultados sobre las bases de datos del repositorio Keel (clasificación multiclase). Las medidas $\overline{Acc}$ y el FM_μ son considerados. La media ± desviación estándar son mostradas para la validación cruzada anidada de 10-fold.

Datasets	SVM OvR	SVM OvO	TBSVM OvR	TBSVM OvO	ETWSVM OvR	ETWSVM OvO	ETWSVM* OvR	ETWSVM* OvO	ETWSVM† OvR	ETWSVM† OvO	ETWSVM†* OvR	ETWSVM†* OvO
Balance	2, 2±0, 1	2, 0±0, 1	65, 4±1, 9	56, 3±1, 4	49, 4±0, 5	46, 5±0, 7	19, 3±1, 2	9, 4±0, 2	0, 8±0, 0	**0, 7±0, 0**	3, 0±0, 1	3, 1±0, 1
Contraceptive	169, 3±44, 3	157, 4±34, 2	500, 2±4, 7	292, 3±4, 9	302, 9±3, 4	202, 6±1, 6	243, 4±38, 7	166, 7±1, 9	4, 5±0, 2	**2, 8±0, 1**	78, 3±1, 9	80, 2±3, 2
Dermatology	3, 3±0, 2	6, 7±0, 0	65, 7±0, 9	83, 0±0, 5	55, 2±1, 0	75, 1±0, 5	15, 0±0, 4	12, 8±0, 2	**0, 8±0, 0**	1, 0±0, 0	3, 2±0, 1	3, 5±0, 2
Ecoli	9, 5±2, 3	15, 8±0, 4	85, 7±1, 2	164, 6±12, 3	72, 2±2, 8	145, 9±1, 4	12, 0±1, 1	15, 0±0, 2	**1, 0±0, 0**	1, 9±0, 1	0, 5±0, 1	0, 6±0, 1
Glass	10, 7±2, 2	7, 7±0, 5	30, 1±0, 3	70, 5±0, 3	25, 7±0, 3	58, 7±0, 2	5, 9±0, 1	7, 3±0, 1	**0, 4±0, 0**	0, 8±0, 0	0, 9±0, 0	0, 9±0, 0
Hayes-roth	1, 5±0, 0	1, 4±0, 0	8, 3±0, 1	13, 9±0, 1	7, 3±0, 0	11, 9±0, 0	3, 0±0, 1	2, 2±0, 1	**0, 1±0, 0**	0, 2±0, 0	0, 6±0, 0	0, 6±0, 0
New-thyroid	1, 8±0, 3	1, 3±0, 0	13, 4±0, 1	18, 4±0, 4	11, 3±0, 1	15, 6±0, 1	4, 0±0, 1	2, 6±0, 1	**0, 2±0, 00**	0, 2±0, 0	0, 7±0, 0	0, 7±0, 0
Iris	1, 7±0, 2	24, 3±0, 6	8, 9±0, 1	445, 0±5, 3	7, 8±0, 0	12, 0±0, 0	3, 2±0, 1	2, 4±0, 1	**0, 1±0, 00**	0, 2±0, 0	0, 7±0, 0	0, 7±0, 0
Thyroid	33, 8±32, 3	24, 2±8, 9	112, 3±1, 3	13, 6±0, 0	81, 9±0, 7	93, 0±1, 1	28, 6±1, 2	10, 1±0, 1	**1, 1±0, 0**	1, 2±0, 0	0, 6±0, 0	0, 6±0, 0
Wine	1, 4±0, 1	12, 5±5, 1	10, 8±0, 1	123, 4±2, 1	9, 4±0, 0	13, 5±0, 0	3, 3±0, 2	2, 9±0, 1	**0, 1±0, 0**	0, 2±0, 0	0, 9±0, 0	0, 9±0, 0
Penbased	152, 7±32, 1	**1, 3±0, 0**	1132, 7±13, 1	15, 5±0, 1	760, 7±9, 8	360, 5±5, 7	180, 5±17, 2	103, 3±1, 8	10, 5±0, 4	4, 8±0, 2	40, 7±0, 8	40, 6±1, 0
Averange	35, 3±63, 0	23, 2±45, 4	184, 9±344, 6	117, 9±137, 5	125, 8±226, 9	94, 1±107, 4	47, 1±83, 1	30, 4±53, 8	1, 8±3, 2	**1, 3±1, 4**	11, 8±25, 1	12, 0±25, 5
Win	0/11	1/11	0/11	0/11	0/11	0/11	0/11	0/11	8/11	2/11	0/11	0/11

Tabla 12.4.: Tiempo de entrenamiento para el repositorio Keel (clasificación multiclase). La media ± la desviación estándar mostradas corresponden al tiempo de entrenamiento y esta dado en segundos por fold.



(a) $\overline{Acc}$ (b) FM_{μ}

Figura 12.7.: Prueba de significancia estadística entre los clasificadores estudiados sobre el repositorio Keel, usando la prueba de Tukey-kramer. Se muestra el diagrama de caja del ranking de media del algoritmo, y el color indica que algoritmos son estadísticamente iguales.

13. Conclusiones

Presentamos un clasificador dedicado a la clasificación de datos desbalanceados nombrado *enhanced twin support vector machine*–(ETWSVM). Nuestra propuesta representa los datos de entrada en un espacio de características de alta dimensionalidad a través de un adecuado mapeo a un RKHS dentro de la optimización del ETWSVM. También, estimamos la función kernel usando una proyección lineal basada en CKA para emparejar las características de entrada con las etiquetas correspondientes. Entonces, el ETWSVM estimula la separabilidad de los datos y usa el conocimiento a priori para sintonizar la función kernel.

En este sentido, nuestro método es una nueva extensión no-lineal que incorpora una formulación dual en un RKHS de dimensión "infinita" para el clasificador TBSVM. Entonces, aplicamos el lemma de la matriz inversa para lograr una estimación basada en datos para los hiperplanos de separación. A su vez, el puntaje de clasificación es calculada como la distancia mínima entre los nuevos datos y los hiperplanos entrenados.

El ETWSVM básico es evaluado dada la configuración de los parámetros de regularización. Luego se estudio la influencia de los parámetros para diferentes escenarios de tasas de desbalance y traslape. Aparte de esto, el algoritmo SMOTE es acoplado con el ETWSVM por el bien de la comparación. Sucesivamente, extendemos el ETWSVM a problemas multiclase por medio de las estrategias OvR y OvO.

Los resultados experimentales se llevan a cabo sobre bases de datos sintéticas y reales. Específicamente, usamos los repositorios UCI y Keel para bases de datos reales para problemas biclase y multiclase, respectivamente. Los resultados logrados en términos del acierto, media geométrica, F-measure y tiempo requerido para el entrenamiento, muestran que nuestro ETWSVM y variantes superan a los métodos del estado del arte. Destacando, al ETWSVM$^{\dagger *}$ como una opción adecuada por su relación de desempeño de clasificación y tiempo de entrenamiento, sin necesidad de usar información a priori sofisticada para sintonizar los parámetros de regularización.

14. Trabajos futuros

Se planea extender el algoritmo ETWSVM para problemas de gran escala [72, 73]. También, probar varios tipos de funciones kernel con su estimación basada en CKA podría ser una linea de investigación interesante [74].

Con respecto al problema multiclase se piensa en la implementación de otras estrategias para extender el ETWSVM y sus variantes a este problema [39].

15. Publicaciones

Como resultado de esta investigación se realizaron las siguientes publicaciones:

Publicaciones en conferencias

- Jimenez C., Diaz D., Salazar D., Alvarez A.M., Orozco A., Henao O. (2018) Nerve Structure Segmentation from Ultrasound Images Using Random Under-Sampling and an SVM Classifier. In: Campilho A., Karray F., ter Haar Romeny B. (eds) Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science, vol 10882. Springer, Cham
- Jimenez C., Alvarez A.M., Orozco A. (2019) A Data Representation Approach to Support Imbalanced Data Classification Based on TWSVM. In: Vera-Rodriguez R., Fierrez J., Morales A. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2018. Lecture Notes in Computer Science, vol 11401. Springer, Cham

Publicaciones en revista

- Jimenez C., Alvarez A.M., Orozco A. (2019), Imbalanced data classification using an enhanced twin support vector machine, In: Pattern Recognition. **(Sometido)**.

Bibliografía

[1] S. Liu, Y. Wang, J. Zhang, C. Chen, and Y. Xiang. Addressing the class imbalance problem in twitter spam detection using ensemble learning. *Computers and Security*, 69:35–49, 2017. cited By 8. 1

[2] M. Bach, A. Werner, J. Żywiec, and W. Pluskiewicz. The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. *Information Sciences*, 384:174–190, 2017. cited By 23. 1

[3] S. Fotouhi, S. Asadi, and M.W. Kattan. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics*, 90, 2019. cited By 1. 1

[4] L. Peng, H. Zhang, Y. Chen, and B. Yang. Imbalanced traffic identification using an imbalanced data gravitation-based classification model. *Computer Communications*, 102:177–189, 2017. cited By 8. 1

[5] J. Sun, J. Lang, H. Fujita, and H. Li. Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates. *Information Sciences*, 425:76–91, 2018. cited By 24. 1

[6] Guo Haixiang and et al. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.*, 73:220–239, 2017. 1, 20, 27

[7] O. Loyola-González, J.F. Martínez-Trinidad, J.A. Carrasco-Ochoa, and M. García-Borroto. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing*, 175:935–947, 2016. cited By 25. 1, 2, 42

[8] P. Branco, L. Torgo, and R.P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2), 2016. cited By 82. 1, 2, 41

[9] Lei Bao, Cao Juan, Jintao Li, and Yongdong Zhang. Boosted near-miss under-sampling on svm ensembles for concept detection in large-scale imbalanced datasets. *Neurocomputing*, 172:198–206, 2016. 1, 3

[10] Maciej Zieba and Jakub M Tomczak. Boosted svm with active learning strategy for imbalanced data. *Soft Computing*, 19(12):3357–3368, 2015. 2, 3

[11] Bernhard Schölkopf and Alexander J Smola. Learning with kernels: Support vector machines, regularization. *Optimization, and Beyond. MIT press*, 1(2), 2002. 2, 6, 8, 33

[12] D. Tomar and S. Agarwal. Twin support vector machine: A review from 2007 to 2014. *Egyptian Informatics Journal*, 16(1):55–69, 2015. cited By 29. 2, 3

[13] Z. Qi, Y. Tian, and Y. Shi. Structural twin support vector machine for classification. *Knowledge-Based Systems*, 43:74–81, 2013. cited By 97. 2

[14] L. Cao and H. Shen. Combining re-sampling with twin support vector machine for imbalanced data classification. pages 325–329, 2017. cited By 3. 2

[15] L. Liu, L. Wang, H. Ji, W. Zang, and D. Li. Between-class discriminant twin support vector machine for imbalanced data classification. volume 2017-January, pages 7117–7122, 2017. cited By 1. 2

[16] Y. Xu, Q. Wang, X. Pang, and Y. Tian. Maximum margin of twin spheres machine with pinball loss for imbalanced data classification. *Applied Intelligence*, 48(1):23–34, 2018. cited By 4. 2

[17] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409-410:17–26, 2017. cited By 42. 2

[18] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. cited By 4583. 2, 42

[19] S. Wojciechowski and S. Wilk. Difficulty factors and preprocessing in imbalanced data sets: An experimental study on artificial data. *Foundations of Computing and Decision Sciences*, 42(2):149–176, 2017. cited By 4. 2

[20] Junru Lu, Chunkai Zhang, and Fengxing Shi. A classification method of imbalanced data base on pso algorithm. In *International Conference of Young Computer Scientists, Engineers and Educators*, pages 121–134. Springer, 2016. 2

[21] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. Distributional random oversampling for imbalanced text classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 805–808. ACM, 2016. 2

[22] Fulong Ren, Peng Cao, Wei Li, Dazhe Zhao, and Osmar Zaiane. Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm. *Computerized Medical Imaging and Graphics*, 55:54–67, 2017. 2

[23] Jaesub Yun, Jihyun Ha, and Jong-Seok Lee. Automatic determination of neighborhood size in smote. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, page 100. ACM, 2016. 2

[24] Chun-Wu Yeh, Der-Chiang Li, Liang-Sian Lin, and Tung-I Tsai. A learning approach with under-and over-sampling for imbalanced data sets. In *Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on*, pages 725–729. IEEE, 2016. 2

[25] X. Zhang, S. Ding, and T. Sun. Multi-class lstmsvm based on optimal directed acyclic graph and shuffled frog leaping algorithm. *International Journal of Machine Learning and Cybernetics*, 7(2):241–251, 2016. cited By 18. 2, 37, 39

[26] W. Deng, L. Deng, J. Liu, and J. Qi. Sampling method based on improved c4.5 decision tree and its application in prediction of telecom customer churn. *International Journal of Information Technology and Management*, 18(1):93–109, 2019. cited By 0. 2

[27] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156, 1996. 3

[28] X. Tao, Q. Li, W. Guo, C. Ren, C. Li, R. Liu, and J. Zou. Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. *Information Sciences*, 487:31–56, 2019. cited By 1. 3

[29] M. Schlögl, R. Stütz, G. Laaha, and M. Melcher. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. *Accident Analysis and Prevention*, 127:134–149, 2019. cited By 0. 3

[30] Yuan-Hai Shao, Wei-Jie Chen, Jing-Jing Zhang, Zhen Wang, and Nai-Yang Deng. An efficient weighted lagrangian twin support vector machine for imbalanced data classification. *Pattern Recognition*, 47(9):3158 – 3167, 2014. 3, 27, 40, 42

[31] Fanyong Cheng, Jing Zhang, and Cuihong Wen. Cost-sensitive large margin distribution machine for classification of imbalanced data. *Pattern Recognition Letters*, 80:107–112, 2016. 3

[32] Gerardo Casañola-Martin, Teresa Garrigues, Marival Bermejo, Isabel González-Álvarez, Nam Nguyen-Hai, Miguel Ángel Cabrera-Pérez, Huong Le-Thi-Thu, et al. Exploring different strategies for imbalanced adme data problem: case study on caco-2 permeability modeling. *Molecular diversity*, 20(1):93–109, 2016. 3

[33] Sara Del Río, Victoria López, José Manuel Benítez, and Francisco Herrera. On the use of mapreduce for imbalanced big data using random forest. *Information Sciences*, 285:112–137, 2014. 3

[34] V. López, S. Del Río, J.M. Benítez, and F. Herrera. Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data. *Fuzzy Sets and Systems*, 258:5–38, 2015. cited By 102. 3

[35] Sang-Hoon Oh. Error back-propagation algorithm for classification of imbalanced data. *Neurocomputing*, 74(6):1058–1061, 2011. 3

[36] D. Tomar and S. Agarwal. A comparison on multi-class classification methods based on least squares twin support vector machine. *Knowledge-Based Systems*, 81:131–147, 2015. cited By 65. 3, 39

[37] A.J. Brockmeier, J.S. Choi, E.G. Kriminger, J.T. Francis, and J.C. Principe. Neural decoding with kernel-based metric learning. *Neural Computation*, 26(6):1080–1107, 2014. cited By 18. 3, 37

[38] A.M. Alvarez-Meza, A. Orozco-Gutierrez, and G. Castellanos-Dominguez. Kernel-based relevance analysis with enhanced interpretability for detection of brain activity patterns. *Frontiers in Neuroscience*, 11(OCT), 2017. cited By 4. 3, 36, 40

[39] S. Ding, X. Zhao, J. Zhang, X. Zhang, and Y. Xue. A review on multi-class twsvm. *Artificial Intelligence Review*, pages 1–27, 2017. cited By 1; Article in Press. 3, 33, 38, 42, 55

[40] A. Roy, R.M.O. Cruz, R. Sabourin, and G.D.C. Cavalcanti. A study on combining dynamic selection and data preprocessing for imbalance learning. *Neurocomputing*, 286:179–192, 2018. cited By 8. 3

[41] P. Xanthopoulos and T. Razzaghi. A weighted support vector machine method for control chart pattern recognition. *Computers and Industrial Engineering*, 70(1):134–149, 2014. cited By 42. 3

[42] S. Piri, D. Delen, and T. Liu. A synthetic informative minority over-sampling (simo) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support Systems*, 106:15–29, 2018. cited By 8. 3

[43] Emanuel Parzen. Statistical inference on time series by hilbert space methods, i. Technical report, STANFORD UNIV CA APPLIED MATHEMATICS AND STATISTICS LABS, 1959. 6

[44] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950. 6

[45] Erwin Kreyszig. *Introductory functional analysis with applications*, volume 1. wiley New York, 1978. 6

[46] George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971. 8

[47] William Karush. Minima of functions of several variables with inequalities as side constraints. Master's thesis, Univ. of Chicago, 1939. 13

[48] Harold W Kuhn and Albert W Tucker. Nonlinear programming. In Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability, 1951. 13

[49] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, INC, 2012. 14

[50] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 14

[51] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995. 14

[52] Yu-Xin Li, Yuan-Hai Shao, and Nai-Yang Deng. Improved prediction of palmitoylation sites using pwms and svm. *Protein and peptide letters*, 18(2):186–193, 2011. 15

[53] Dino Isa, Lam H Lee, VP Kallimani, and Rajprasad Rajkumar. Text document preprocessing with the bayes formula for classification using the support vector machine. *IEEE Transactions on Knowledge and Data engineering*, 20(9):1264–1272, 2008. 15

[54] Yingjie Tian, Yong Shi, and Xiaohui Liu. Recent advances on support vector machines research. *Technological and Economic Development of Economy*, 18(1):5–33, 2012. 15

[55] Edgar Elias Osuna. *Support vector machines: Training and applications*. PhD thesis, Massachusetts Institute of Technology, 1998. 20

[56] Olvi L Mangasarian and Edward W Wild. Multisurface proximal support vector machine classification via generalized eigenvalues. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):69–74, 2006. 21

[57] and R. Khemchandani and Chandra Suresh. Twin support vector machines for pattern classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):905–910, May 2007. 22, 24, 25

[58] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. 1998. 24

[59] Y. Shao, C. Zhang, X. Wang, and N. Deng. Improvements on twin support vector machines. *IEEE Transactions on Neural Networks*, 22(6):962–968, June 2011. 26, 42

[60] C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13:795–828, 2012. cited By 151. 36

[61] J. López, S. Maldonado, and M. Carrasco. A robust formulation for twin multiclass support vector machine. *Applied Intelligence*, 47(4):1031–1043, 2017. cited By 1. 38

[62] Y.-H. Shao, W.-J. Chen, Z. Wang, C.-N. Li, and N.-Y. Deng. Weighted linear loss twin support vector machine for large-scale classification. *Knowledge-Based Systems*, 73(1):276–288, 2014. cited By 38. 39

[63] X. Hua and S. Ding. Weighted least squares projection twin support vector machines with local information. *Neurocomputing*, 160:228–237, 2015. cited By 15. 39

[64] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009. cited By 1109. 41

[65] X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(2):539–550, 2009. cited By 614. 42

[66] X. Zhang, Y. Liang, J. Zhou, and Y. Zang. A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimized svm. *Measurement: Journal of the International Measurement Confederation*, 69:164–179, 2015. cited By 163. 42

[67] B.M. Abidine, L. Fergani, B. Fergani, and M. Oussalah. The joint use of sequence features combination and modified weighted svm for improving daily activity recognition. *Pattern Analysis and Applications*, 21(1):119–138, 2018. cited By 6. 42

[68] J. Mathew, C.K. Pang, M. Luo, and W.H. Leong. Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, 29(9):4065–4076, 2018. cited By 7. 42

[69] José Alberto Hernández-Muriel, Andrés Marino Álvarez-Meza, Julián David Echeverry-Correa, Álvaro Ángel Orozco-Gutierrez, and Mauricio Alexánder Álvarez-López. Feature relevance estimation for vibration-based condition monitoring of an internal combustion engine. *Tecno Lógicas*, 20(39):159–174, 2017. 48

[70] J. Pizarro, E. Guerrero, and P.L. Galindo. Multiple comparison procedures applied to model selection. *Neurocomputing*, 48(1-4):155–173, 2002. cited By 40. 51

[71] H.D. Vargas Cardona, A.A. Orozco, and M.A. Álvarez. Multi-patient learning increases accuracy for subthalamic nucleus identification in deep brain stimulation. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, pages 4341–4344, 2012. cited By 1. 51

[72] M. Goto, R. Ishida, and S. Uchida. A preselection-based fast support vector machine learning for large-scale pattern sets using compressed relative neighborhood graph. *Research Reports on Information Science and Electrical Engineering of Kyushu University*, 22(1):1–7, 2017. cited By 0. 55

[73] S. Sharma and R. Rastogi. Stochastic conjugate gradient descent twin support vector machine for large scale pattern classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11320 LNAI:590–602, 2018. cited By 0. 55

[74] M. Zareapoor, P. Shamsolmoali, D. Kumar Jain, H. Wang, and J. Yang. Kernelized support vector machine with deep learning: An efficient approach for extreme multiclass dataset. *Pattern Recognition Letters*, 115:4–13, 2018. cited By 4. 55